

CHEP 2015 - Highlights

Oliver Gutsche, with input from Philippe Canal

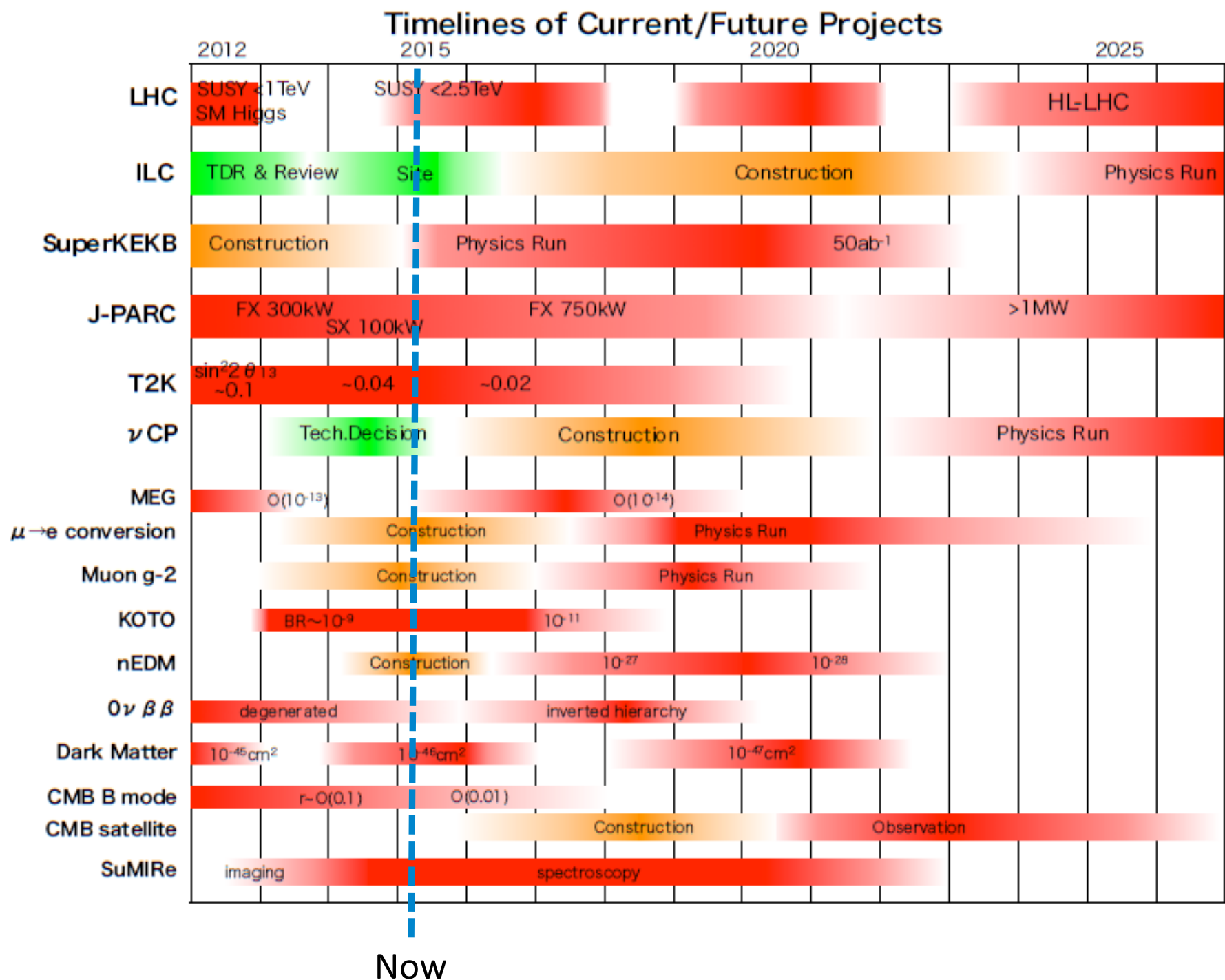
After-Lunch-Seminar at Fermilab

21. May 2015

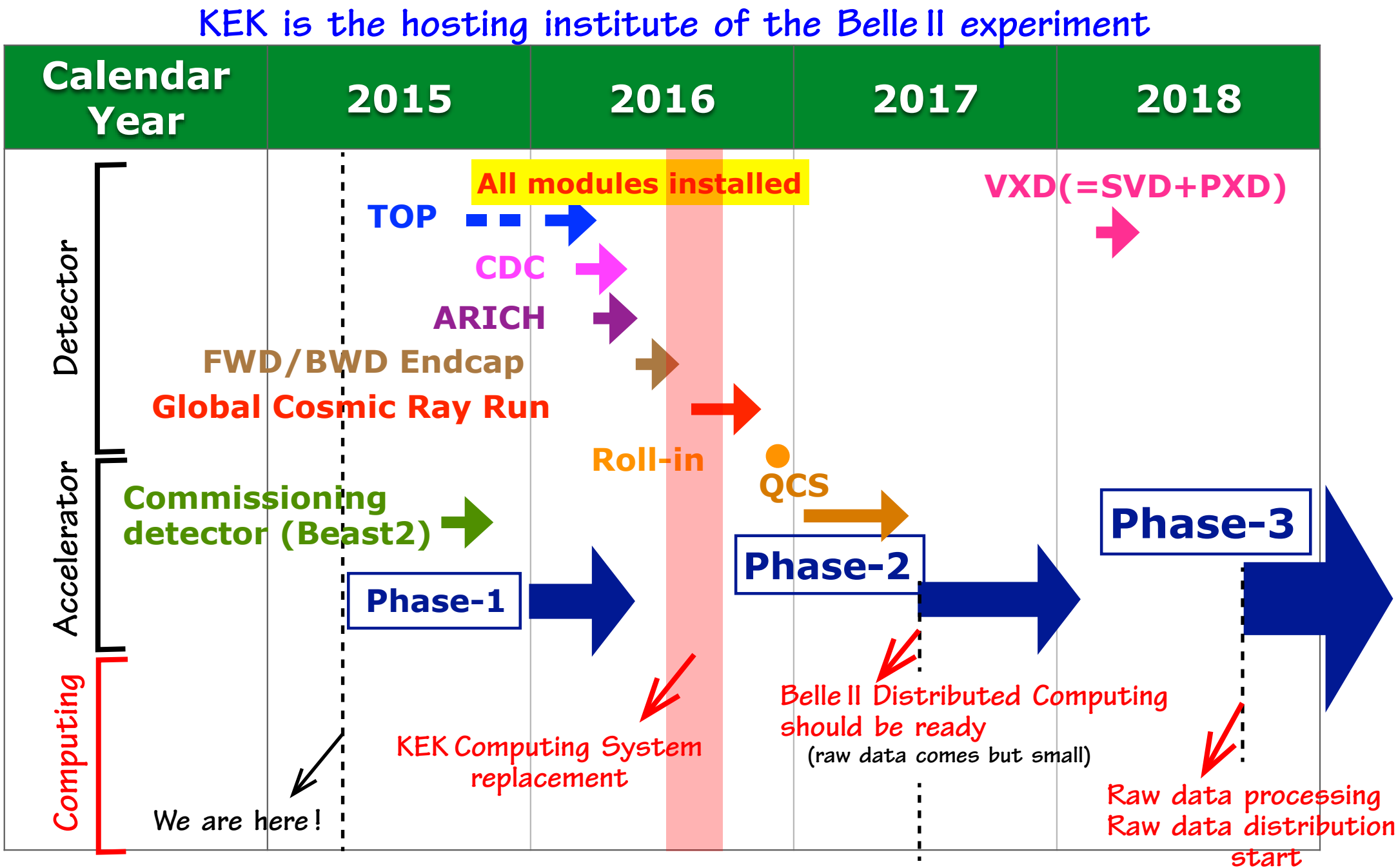
Future plan of KEK & Belle II



Roadmap of Japanese HEP community



SuperKEKB/Belle II Time line



- <https://indico.cern.ch/event/304944/session/15/contribution/549/material/slides/0.pptx>
- <https://indico.cern.ch/event/304944/session/15/contribution/550/material/slides/0.pdf>

Technology: Baseline
Boundary Conditions in 2025

Technology	Growth in 10 years
CPU Servers	x4 - 14
Disk Capacity	x4 - 10
Tape Capacity	x10 - 30
Network Capacity	x30 - 200

- Per unit cost
- Assumes no truly radical change in what we do and no massively disruptive technological advances

13

Future Processor Directions:
General vs Specialist Trends

- At the same time applications which can be considered 'bulk-specialist' are cost effective to tackle with a custom architecture (lower power is the 'killer feature')
 - FPGAs could be used to run cloud apps such as voice recognition and search (with OpenCL as a programming interface)
 - See Intel and Altera's collaboration on HARP
 - 12-core Intel microprocessor with an Altera Stratix V FPGA module
 - Specifically targeting the development of tools to broaden the base of suitable applications
- N.B. we do not have to migrate the entire workload — just enough to fill the available hardware resource (evgen, sim)
 - Note that attached to HPCs we often have exotic hardware that might sit idle

Summary: processor futures look very heterogenous and very interesting - lots of opportunities for smart work.
(Did someone mention validation of the results and physics performance...?)



21

▪ <https://indico.cern.ch/event/304944/session/15/contribution/551/material/slides/0.pdf>

Plenary: Evolution of Computing and Software at LHC: from Run 2 to HL-LHC

Estimating the Computing Challenge (for General Purpose Detectors)

- To make an estimate of how much computing we need at LHC we have to understand the scaling between $\mu=40$ and $\mu=140$ and HLT output rates of 1kHz vs 5kHz

Step	HL-LHC do nothing factor	HL-LHC smart factor	Comment
Generation	x20	x5	Need more precise and heavily filtered generators
Simulation	x5	x3	Does not scale with μ , but we will need more simulation; 'smart' includes GEANT improvements and fast sim
Digitisation	x20	x10	Linear with μ , technical improvements possible
Reco (MC)	x100	x15	Non-linear with μ , plus need more events; smart includes truth tracking and algorithmic improvements
Reco (Data)	x100	x25	Non-linear with μ , plus need more events; smart includes algorithmic improvements and maybe track trigger info
Analysis	x10	x5	Scales mostly with data volume, main problem probably i/o

Personal
guesstimates

45

GPD Processing Challenge at High Luminosity

Step	Approx. Fraction Today	HL-LHC do nothing multiplication factor	HL-LHC do nothing CPU increase	HL-LHC smart multiplication factor	HL-LHC smart CPU increase
Generation	0.05	20	1	5	0.25
Simulation	0.45	5	2.25	3	1.35
Digitisation	0.05	20	1	10	0.5
Reco (MC)	0.15	100	15	15	2.25
Reco (Data)	0.1	100	10	25	2.5
Analysis	0.2	10	2	5	1
Total (in units of today's compute)	1		31.25		7.85

- This is a straw man model, but I really believe that
 - Do nothing is not an option (technically as well as politically)
 - Smart gets us into the domain of the possible

46

▪ <https://indico.cern.ch/event/304944/session/15/contribution/551/material/slides/0.pdf>

Plenary: Computing in Intensity Frontier Accelerator Experiments



Small Experiments?

- Intensity Frontier experiments are not small!



April, 2015

C. Group - UVA and Fermilab

14



Common Services and Projects Toolkit

FIFE (Fabric for Frontier Experiments): Manages/provides access to shared resources at Fermilab
Used by all neutrino and muon experiments at FNAL

FIFE provides access and support for a comprehensive set of services and tools:

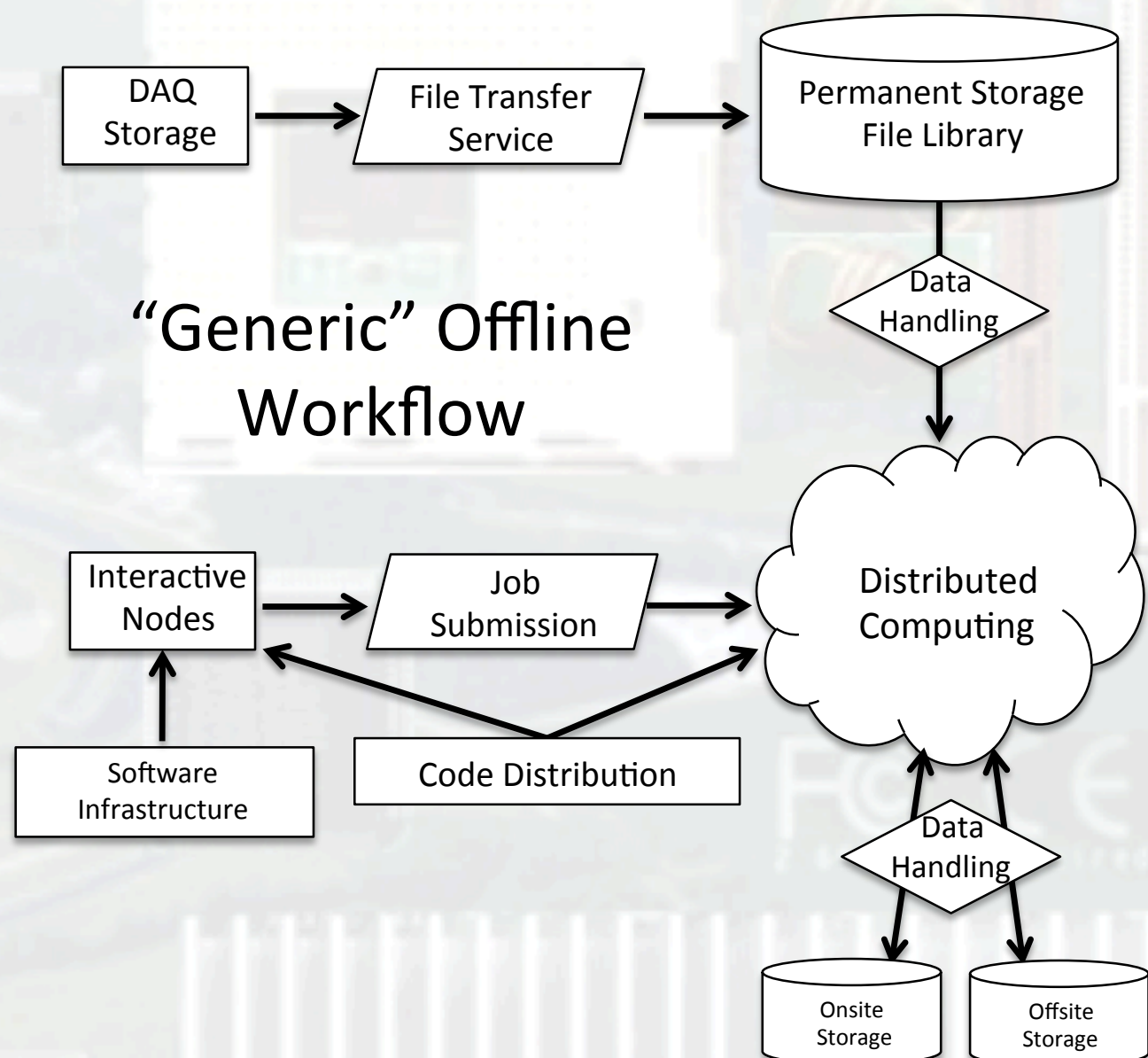
- DAQ and Controls
- Grid and Cloud
- Scientific Data Storage and Access
- Scientific Data Management
- Scientific Frameworks and Software
- Physics and detector simulation
- Databases
- Scientific Computing Systems
- Scientific Collaboration Tools

See talk on Thursday, Parag Mhashilkar, Track 4:

Advances in Distributed High Throughput Computing for the Fabric for Frontier Experiments Project at Fermilab

April, 2015

C. Group - UVA and Fermilab



Slide content, O. Gutsche

▪ <https://indico.cern.ch/event/304944/session/15/contribution/554/material/slides/0.pdf>

Plenary: Challenges of Developing and Maintaining HEP "Community" Software

Community Tools: Grid Compute Projects

SLAC

- To support the LHC distributed model, grid computing projects intended to support multiple experiments and disciplines
- WLCG-EGEE & Open Science Grid
 - » Conceived roughly 1999 with funding around 2002
 - » Largely targeted towards access to large scale compute resources
 - » Pushed boundaries for the better: cyber security
- Challenges
 - » Maintaining Funding: Soft money/essential infrastructure.
 - Continual re-invention and restructuring required
 - » Maintaining technical currency (See additional talks)
 - » Communities beyond LHC



13

Recap of the general lessons

SLAC

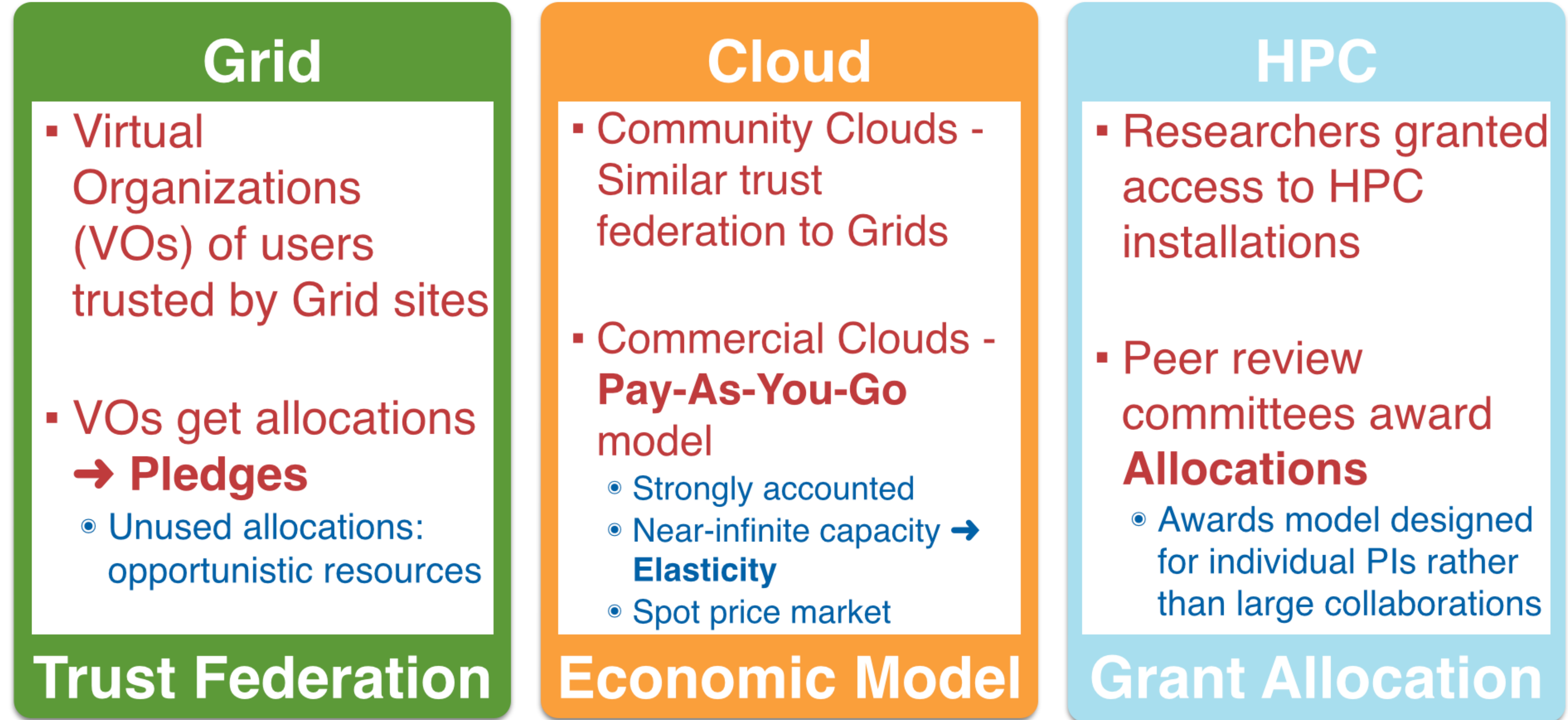
- Note the time scales
 - » Prototyping to production is typically 10 years
- Many different governance models can be effective
- Resources are always limited
 - » Too few people doing too much
 - » Never enough hardware
 - » Increasing scrutiny about 'commercial' solutions
- Staying connected to the community and its needs
 - » Inattention to community software has led to gaps in the portfolio
 - » Burgeoning needs in the neutrino community/Direct Dark Matter
- Providing intellectual and technical continuity
- Managing the projects through the full life cycle
 - » Enabling R&D
 - » Fostering innovation—what are the next great ideas?
 - » Where can and should we move on?

CHEP15--OIST

19

- <https://indico.cern.ch/event/304944/session/15/contribution/556/material/slides/0.pdf>

Plenary: Diversity in Computing Technologies and Strategies for Dynamic Resource Allocation



- <https://indico.cern.ch/event/304944/session/15/contribution/559/material/slides/0.pdf>

Plenary: Distributed Data Management and Distributed File Systems

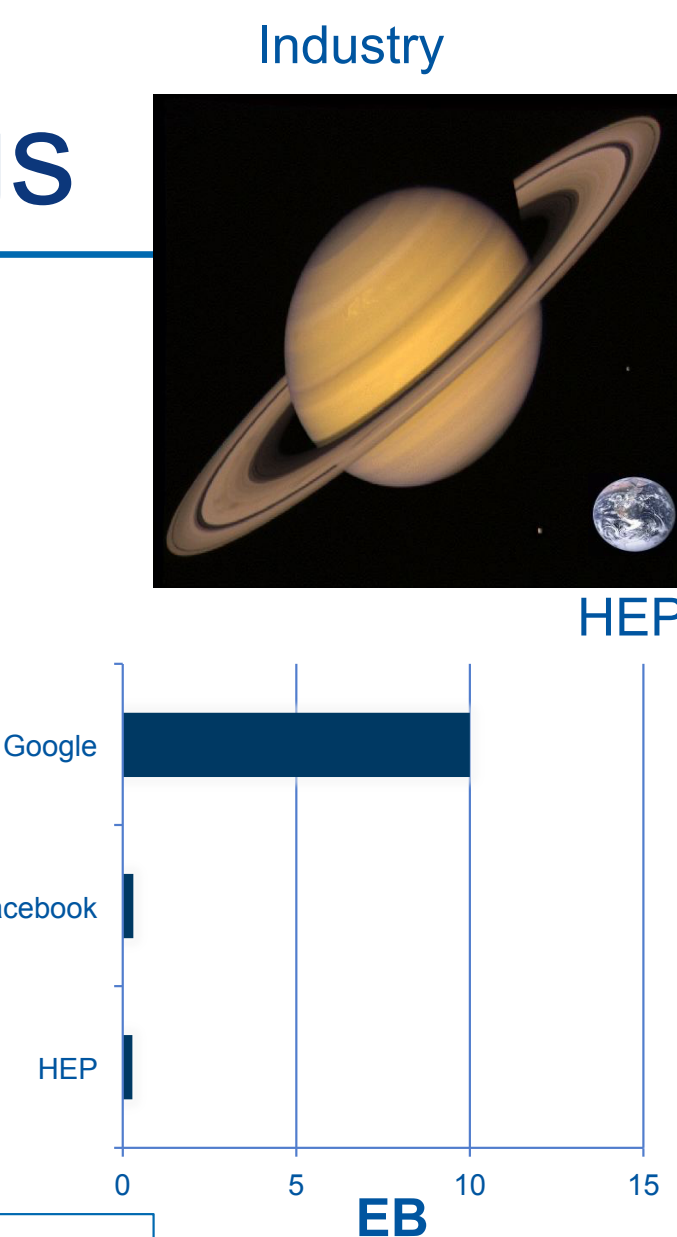
Relative Size of Things

Processing

- Amazon has more than 40 million processor cores in EC2

Storage

- Amazon has 2×10^{12} unique user objects and supports 2M queries per second
- Google has 10-15 exabytes under management
- Facebook 300PB
- eBay collected and accessed the same amount of data as LHC Run1



Our data and processing problems are ~1% the size of the largest industry problems

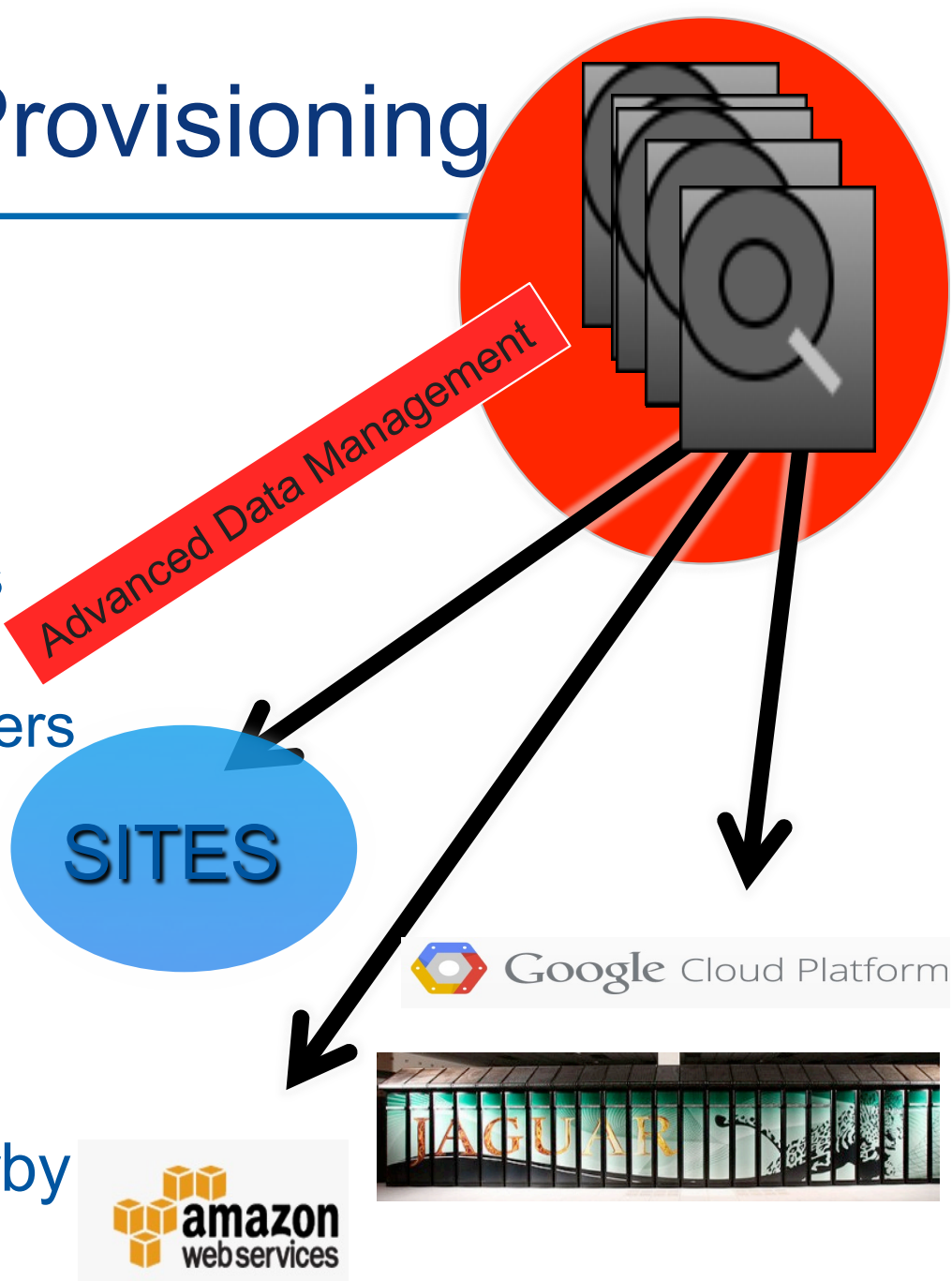


Maria Girone, CHEP 2015

21

A Model for Resource Provisioning

- In this model HEP owns and operates the **Advanced Data Management**
 - Owns storage, provisions network, manages and delivers data
 - Creating large virtual data centers
- The **processing** is **separate, dynamic, and generic**
 - Could be commodity systems, provided by a cloud provider, or a super computer
- **Caching** provides **volatile** nearby smart storage services

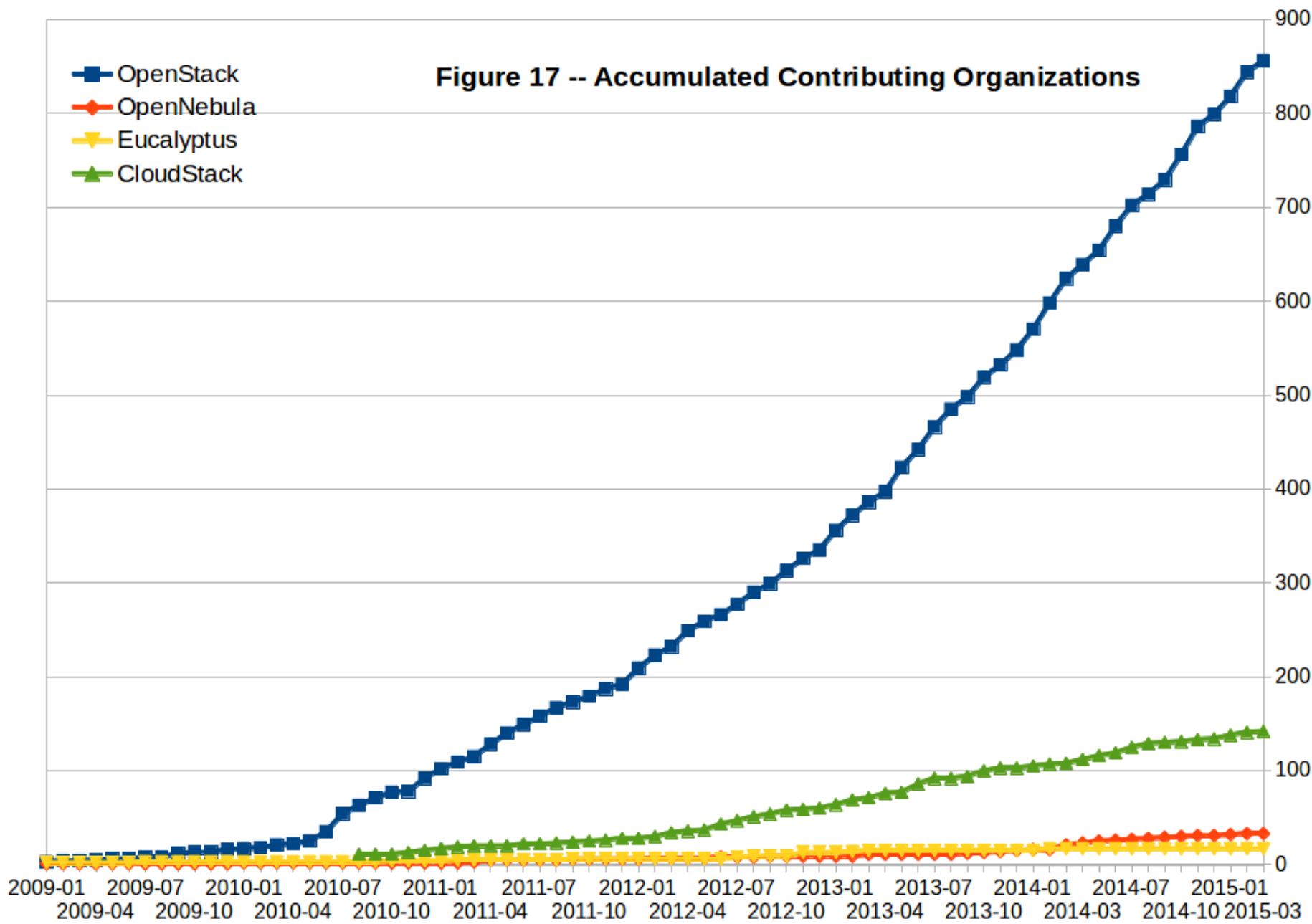


Maria Girone, CHEP 2015

37

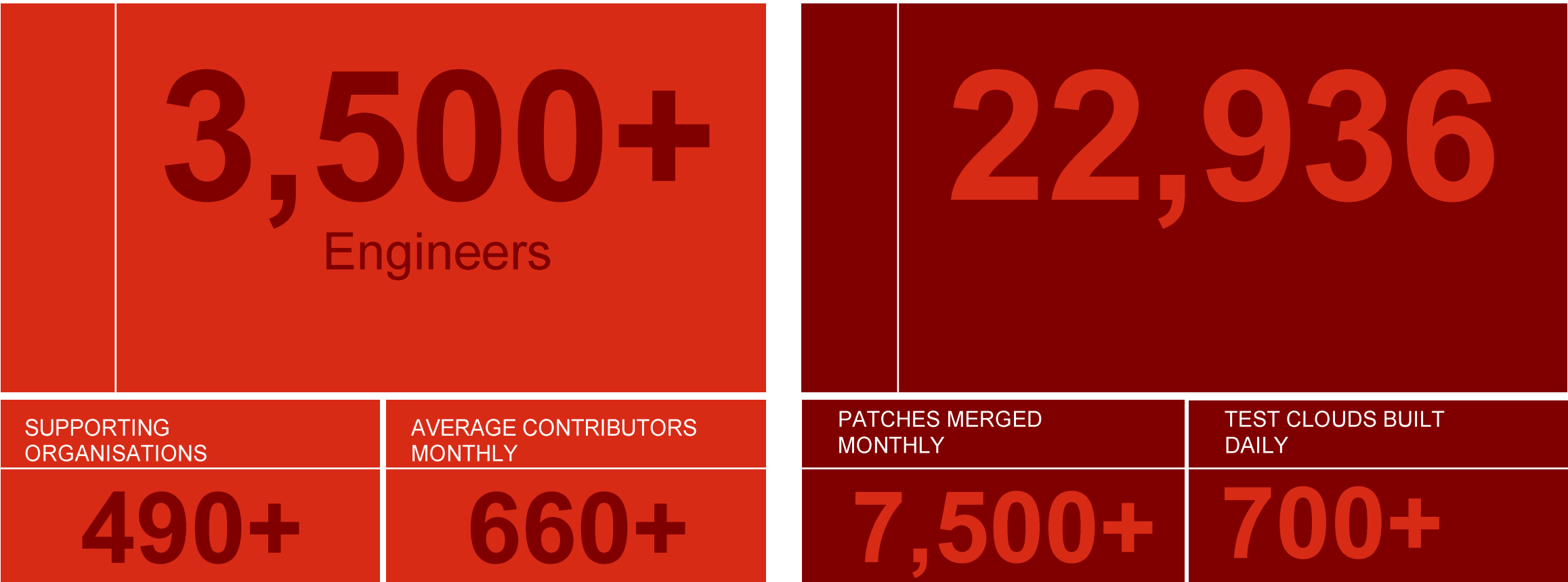
▪ <https://indico.cern.ch/event/304944/session/15/contribution/560/material/slides/1.pptx>

Plenary: Openstack



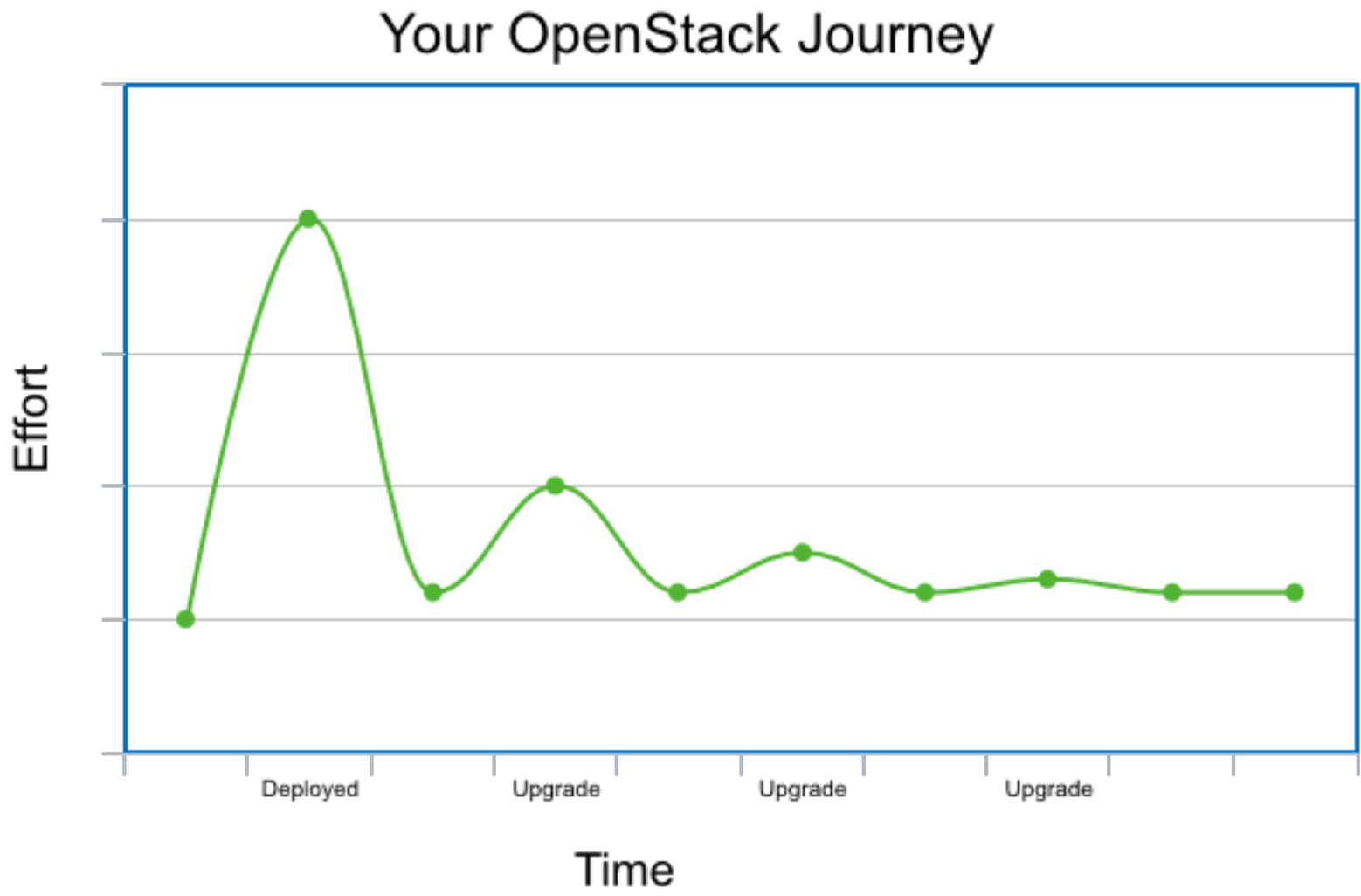
<http://www.udwork.com/item/14057.html> - Qingye Jiang (John), University of Sydney, May 2015

Project is very active with many contributors



- <https://indico.cern.ch/event/304944/session/15/contribution/562/material/slides/0.pdf>

9



30

Plenary: security

DRAM *rowhammer* bug => kernel exploit

Access repeatedly a row of DRAM memory

```
code1a:
    mov (X), %eax    // Read from address X
    mov (Y), %ebx    // Read from address Y
    clflush (X)      // Flush cache for address X
    clflush (Y)      // Flush cache for address Y
    jmp code1a
```

This can cause bit flips in neighboring rows

Proof-of-concepts: **privilege escalation exploits**

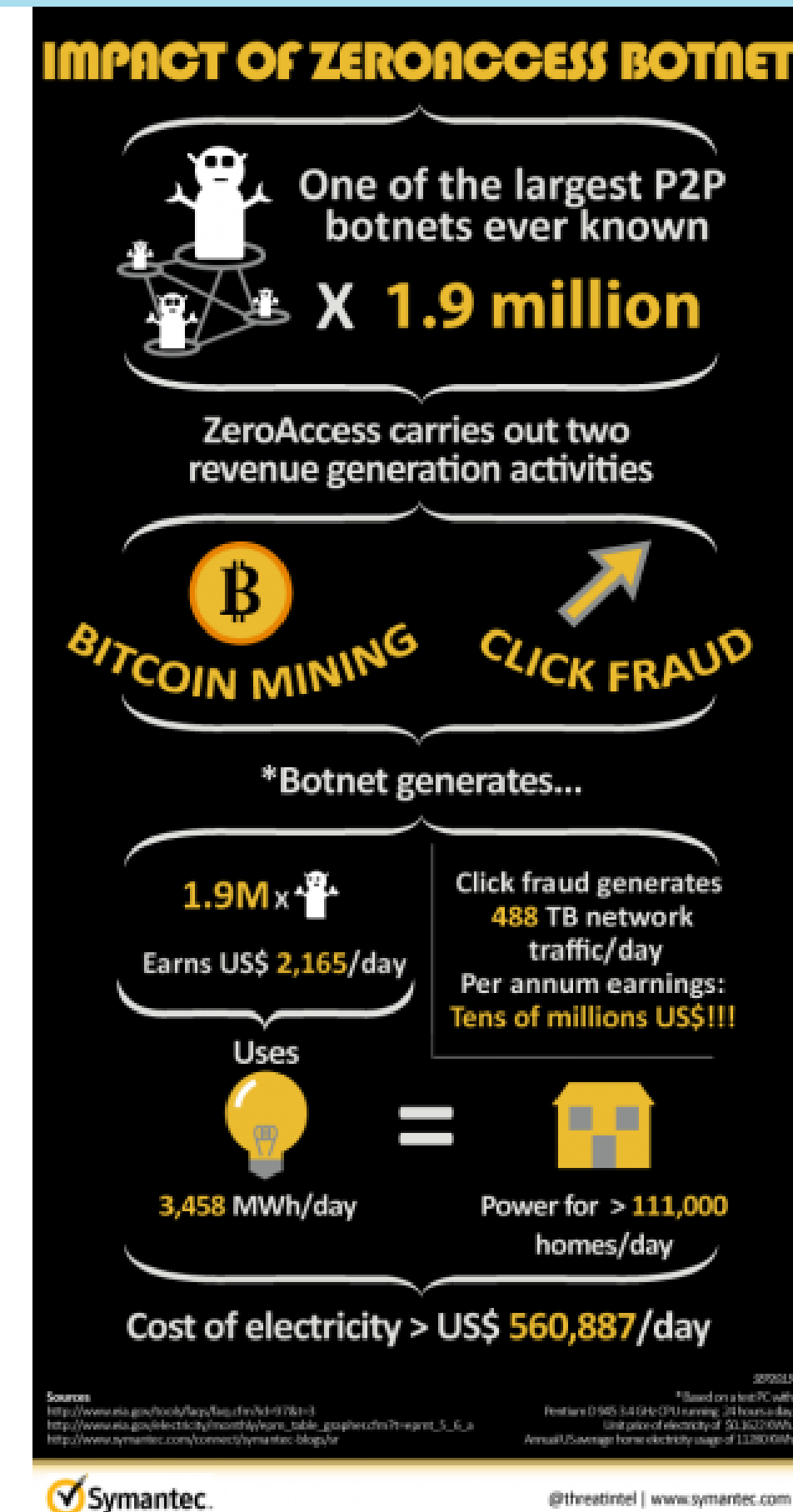
- e.g. modifying page table entries (PTEs)
- goal: gain write access to its own page table
- result: **gain read-write access to all of physical memory**

<http://googleprojectzero.blogspot.fr/2015/03/exploiting-dram-rowhammer-bug-to-gain.html>

- <https://indico.cern.ch/event/304944/session/15/contribution/563/material/slides/1.pdf>

Bitcoin mining:
\$250 K / month

Click fraud:
\$2.8 M / month



16

22

8 months

“Average time between intrusion and detection”

Two types of organisations:
those that know they've been hacked
and those that don't know

Plenary: Networking

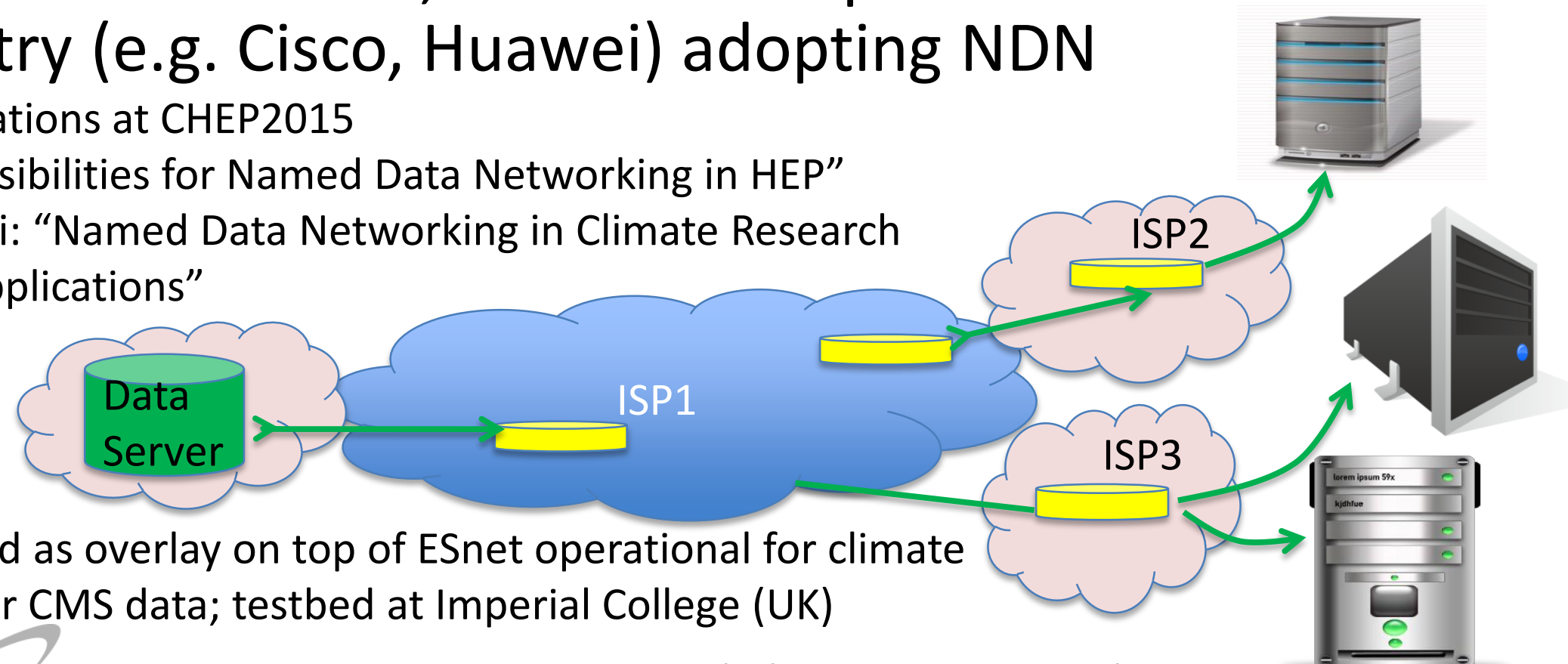
Example: Content Delivery with Named Data Network (NDN)

- Network uses application data names for delivery
 - Multiple users request the same data: network can retrieve from nearby copy
 - Provides performance estimates (user provided metrics)
- Name + data-signature enables in-network storage (sec.)
- Caching happens automatically
- Broadcast “interest”, location-independent data retrieval
- Industry (e.g. Cisco, Huawei) adopting NDN

NDN presentations at CHEP2015

D. Rand: “Possibilities for Named Data Networking in HEP”

S. Shannigrahi: “Named Data Networking in Climate Research and HEP Applications”



BROOKHAVEN

Michael Ernst, BNL CHEP 2015, Okinawa

49

BROOKHAVEN

Michael Ernst, BNL CHEP 2015, Okinawa

50

Conclusions

- Regardless of the resource composition (distributed grid centers, consolidation within a few large data centers, consolidation within clouds, NDN) - high performance networking will continue to be critical to HEP
 - The network as a partner motivates why we should worry about networks
 - Regardless of Computing Models, HEP and network partners will need to work closely together to build the intent-based interfaces between applications and networks that can most effectively accelerate discovery
- Excellent networks, flexible and adaptable computing models and software systems are the foundation to fully exploiting resources such as Grids, Clouds and HPCs
 - Networks overcome limitations of geography
 - To optimize usage of excellent network infrastructure we have access to we need to interact with the control plane in an intelligent way
 - Network Virtualization - integration of storage, compute and network - in a seamless manner, including cloud and local resources. Leveraging efforts like OpenStack to instantiate VMs, allocate storage, and network dynamically
- Named-Data Networking – a new way of accessing content than worrying about where the data is located.

▪ <https://indico.cern.ch/event/304944/session/15/contribution/566/material/slides/2.pdf>

Summary: Track 7 Clouds & Virtualization

Themes

- Emphasis on running jobs
 - Rather than coverage of virtualization as a platform for services
- VM-based jobs systems are used for routine production work
 - CloudScheduler, GlideinWMS, and the 3 Vacuum-based platforms seem to be biggest numerically
- OpenStack prominent at sites but other IaaS platforms are common too
- Commercial cloud use still almost all relies on (large scale) donated resources from providers
- Volunteer computing with BOINC is rapidly emerging as very large potential source of capacity
- Many talks mentioned that they could or would support containers as an alternative to full-blown VMs
- Container-specific talks mostly emphasised packaged apps paradigm

Track 7 Clouds & Virtualization - Andrew.McNab@cern.ch - CHEP 2015, 16 Apr 2015, Okinawa

6

- <https://indico.cern.ch/event/304944/session/15/contribution/569/material/slides/0.pdf>

- Ulrich presented detailed work on how to classify and benchmark VMs, and then inject performance numbers into the accounting system.
 - <https://indico.cern.ch/event/304944/session/7/contribution/86/material/slides/0.pdf>
- Andrew explained how Containers differ from Virtual Machines, and presented performance comparisons with (unoptimised) VMs, with benchmarks and with realistic HEP workloads.
 - <https://indico.cern.ch/event/304944/session/7/contribution/356/material/slides/0.pdf>
- Miguel explained how they are supporting volunteer computing with BOINC in HEP, enabled by virtualization.
 - <https://indico.cern.ch/event/304944/session/7/contribution/82/material/slides/0.pdf>
- Marek explained how identity federation is now possible between OpenStack sites
 - <https://indico.cern.ch/event/304944/session/7/contribution/226/material/slides/0.pdf>

Batch systems: Two Years of HTCondor at the RAL Tier-1



Ongoing work & future plans

- Integration with private cloud
 - OpenNebula cloud setup at RAL, currently with ~1000 cores
 - Want to ensure any idle capacity is used, so why not run virtualized worker nodes?
 - Want opportunistic usage which doesn't interfere with cloud users
 - Batch system expands into cloud when batch system busy & cloud idle
 - Batch system withdraws from cloud when cloud becomes busy
 - Successfully tested, working on moving this into production
 - See posters at CHEP
- Upgrade worker nodes to SL7
 - Setup SL6 worker node environment in a chroot, run SL6 jobs in the chroot using NAMED_CHROOT functionality in HTCondor
 - Will simplify eventual migration to SL7 – can run both SL6 and SL7 jobs
 - Successfully tested CMS jobs

38



Ongoing work & future plans

- Simplification of worker nodes
 - Testing use of CVMFS grid.cern.ch for grid middleware
 - 540 packages installed vs 1300 for a normal worker node
 - HTCondor can run jobs:
 - In chroots
 - In filesystem namespaces
 - In PID namespaces
 - In memory cgroups
 - In CPU cgroups
 - Do we really need pool accounts on worker nodes?
 - With the above, one job can't see any processes or files associated with any other jobs on the same worker node, even if the same user
 - Worker nodes and CEs could be much simpler without them!

39

▪ virtual facility like functionality (ask Andrew Lahiff, CMS person)

- <http://indico.cern.ch/event/345619/session/0/contribution/15/material/slides/1.pdf>

• Posters:

- Experience with batch systems and clouds sharing the same physical resources: <https://indico.cern.ch/event/304944/session/10/contribution/452>
- Integrating grid and cloud resources at the RAL Tier-1: <https://indico.cern.ch/event/304944/session/9/contribution/449>

WLCG: Commercial Clouds and Vacuum Model

Commercial Clouds

- Helix Nebula
 - A public-private partnership
 - Between research organizations and IT industry
- Microsoft Azure Pilot
 - Preliminary discussions with CERN OpenLab
- Amazon
 - BNL RACF for ATLAS and CMS
 - With new Scientific Computing group at AWS
- Deutsche Börse Cloud Exchange AG
 - Beta testing platform
 - Will go live beginning of May
- PICSE
 - Procurement Innovation for Cloud Services in Europe
- European Science Cloud Pilot
 - Pre-Commercial Procurement (PCP) proposal
 - Buyers group public organizations that are members of the WLCG collaboration

Further simplification: Vacuum model

- Following the CHEP 2013 paper:
 - “The Vacuum model can be defined as a scenario in which virtual machines are created and contextualized for experiments by the resource provider itself. The contextualization procedures are supplied in advance by the experiments and launch clients within the virtual machines to obtain work from the experiments’ central queue of tasks.”
(“Running jobs in the vacuum”, A McNab et al 2014 J. Phys.: Conf. Ser. 513 032065)
 - a loosely coupled, late binding approach in the spirit of pilot frameworks
- For the experiments, VMs appear by “spontaneous production in the vacuum”
 - Like virtual particles in the physical vacuum: they appear, potentially interact, and then disappear
- CernVM-FS and pilot frameworks mean a small user_data file and a small CernVM image is all the site needs to create a VM
 - Experiments can provide a template to create the site-specific user_data

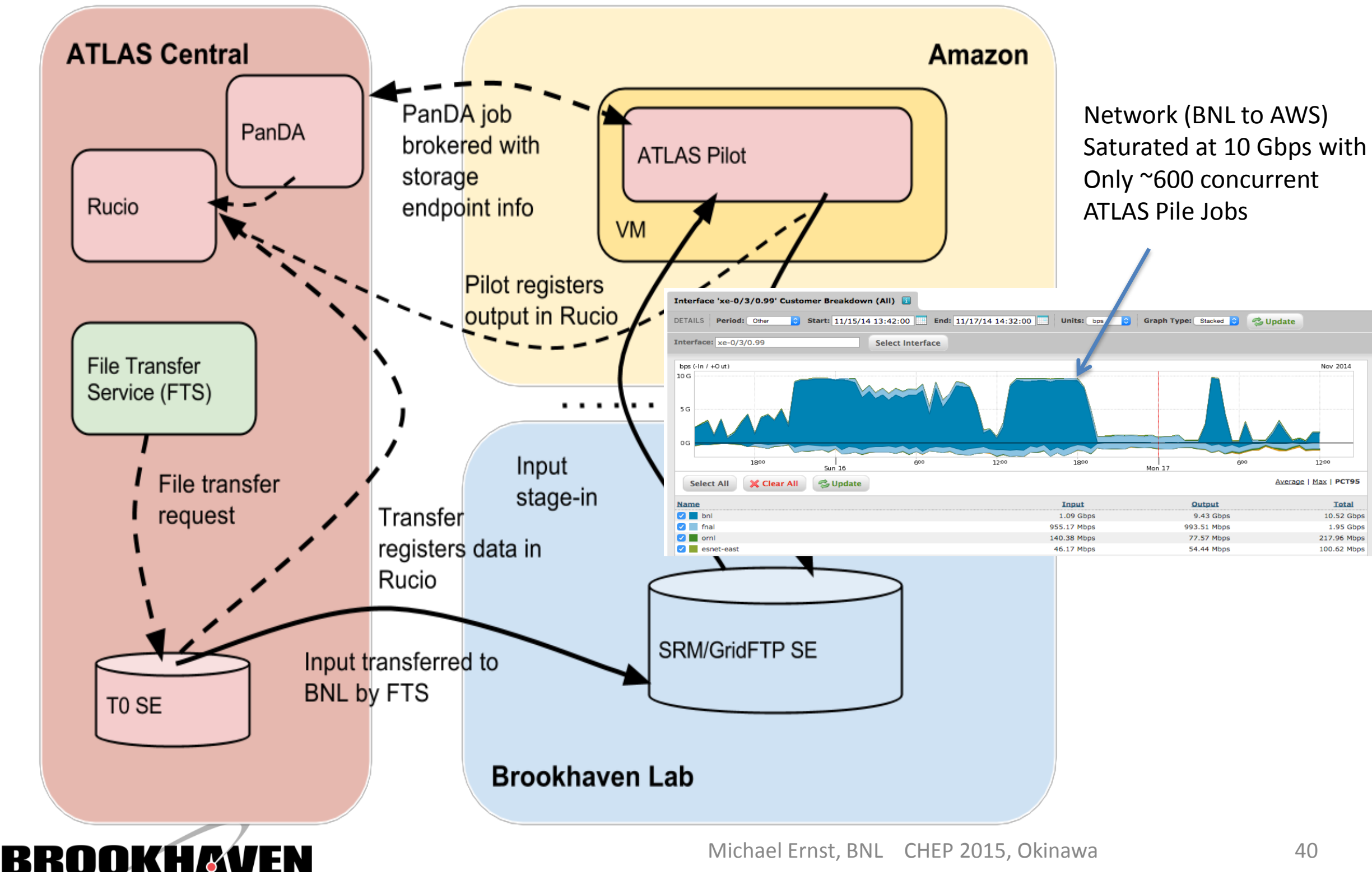
CPU towards 2022 - Andrew.McNab@cern.ch - WLCG Workshop, 12 Apr 2015, Okinawa

▪ <http://indico.cern.ch/event/345619/session/1/contribution/11/0/material/slides/1.pdf>

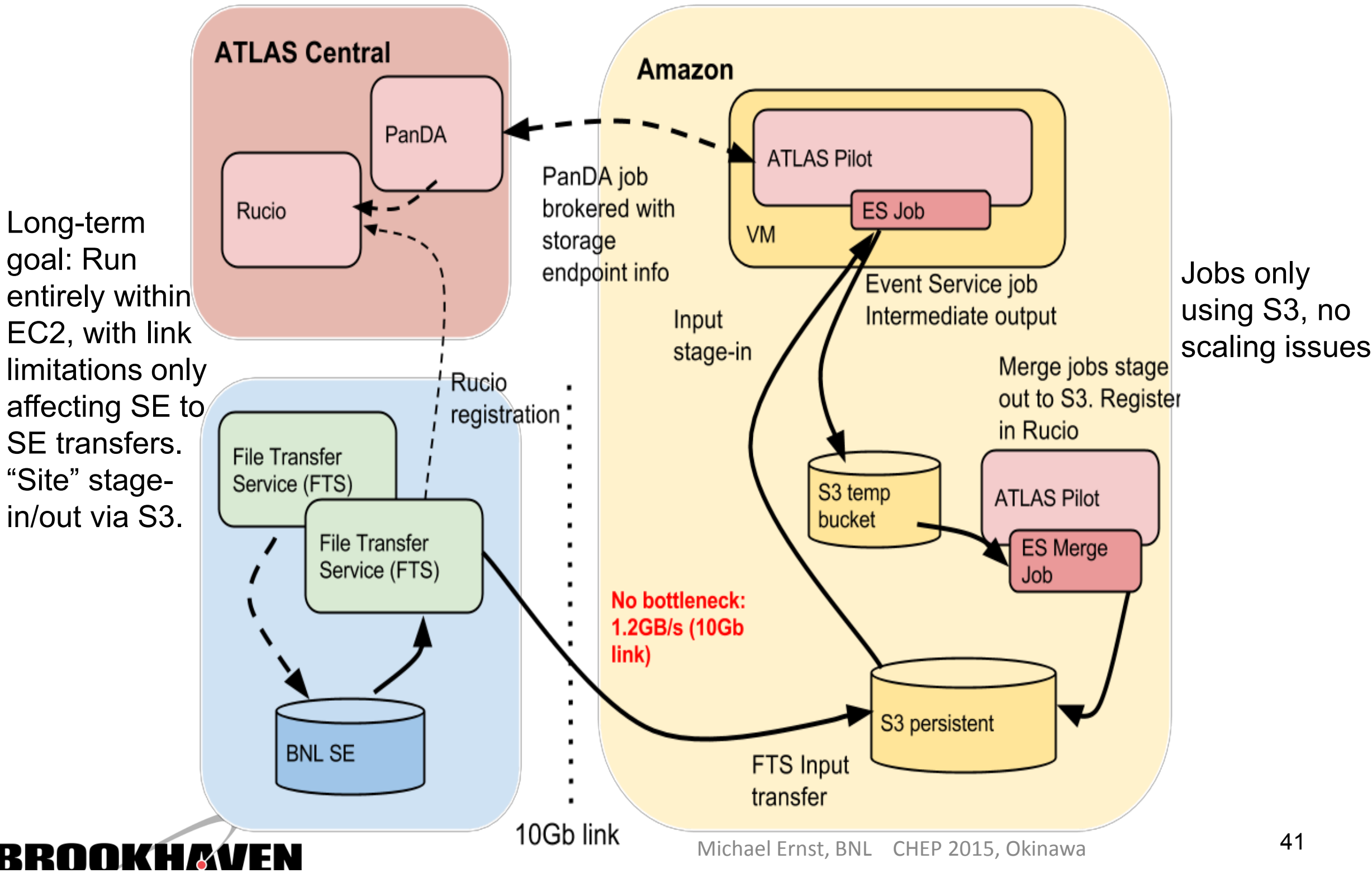
▪ <http://indico.cern.ch/event/345619/session/1/contribution/11/1/material/slides/0.pdf>

CHEP Networking Plenary

Utilize Cloud for Compute ... the initial, easy step to using the “Cloud”



Using the Cloud for Compute and Storage



▪ <https://indico.cern.ch/event/304944/session/15/contribution/566/material/slides/2.pdf>

Summary: Track 7 Clouds & Virtualization

Themes

- Emphasis on running jobs
 - Rather than coverage of virtualization as a platform for services
- VM-based jobs systems are used for routine production work
 - CloudScheduler, GlideinWMS, and the 3 Vacuum-based platforms seem to be biggest numerically
- OpenStack prominent at sites but other IaaS platforms are common too
- Commercial cloud use still almost all relies on (large scale) donated resources from providers
- Volunteer computing with BOINC is rapidly emerging as very large potential source of capacity
- Many talks mentioned that they could or would support containers as an alternative to full-blown VMs
- Container-specific talks mostly emphasised packaged apps paradigm

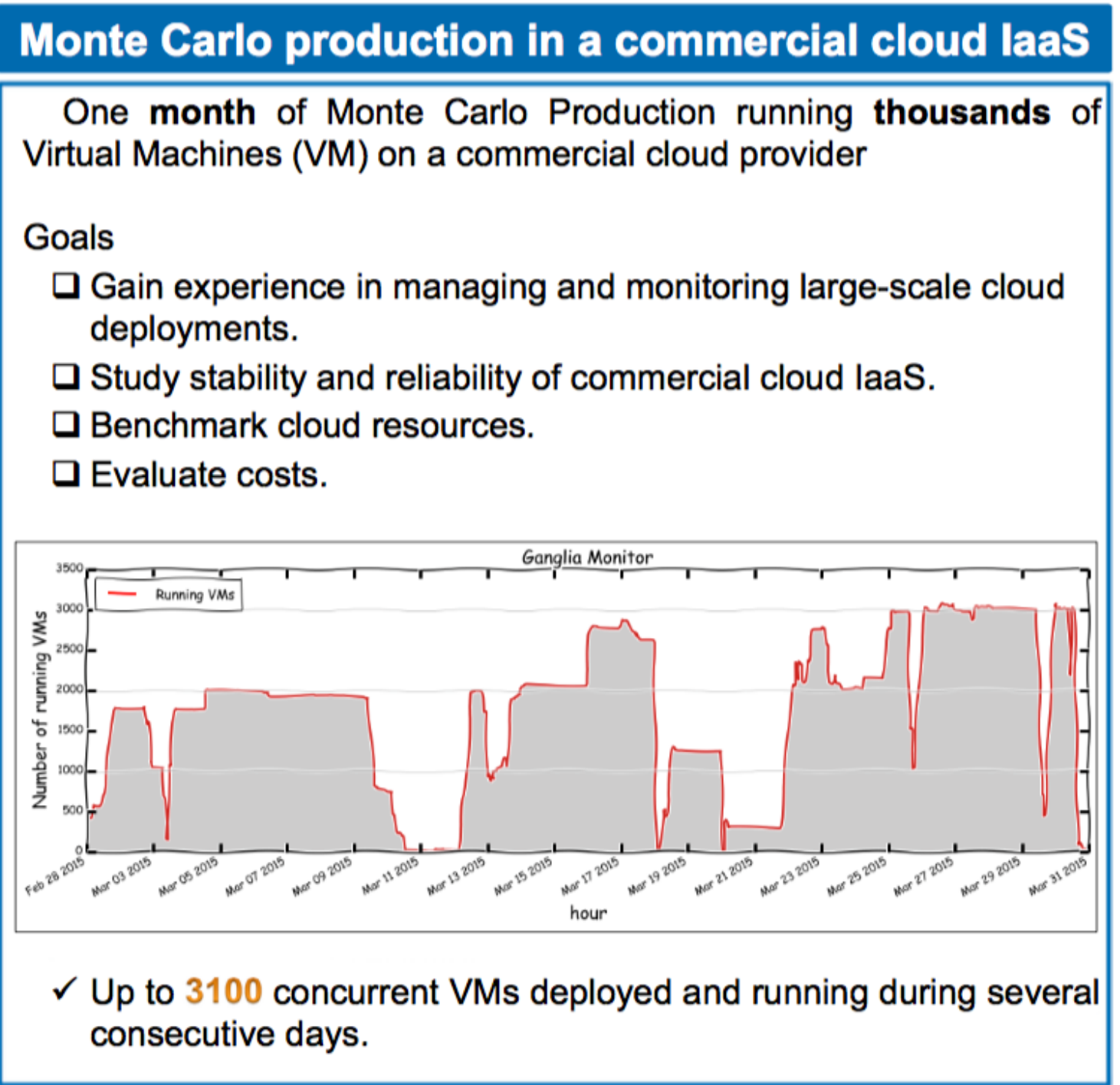
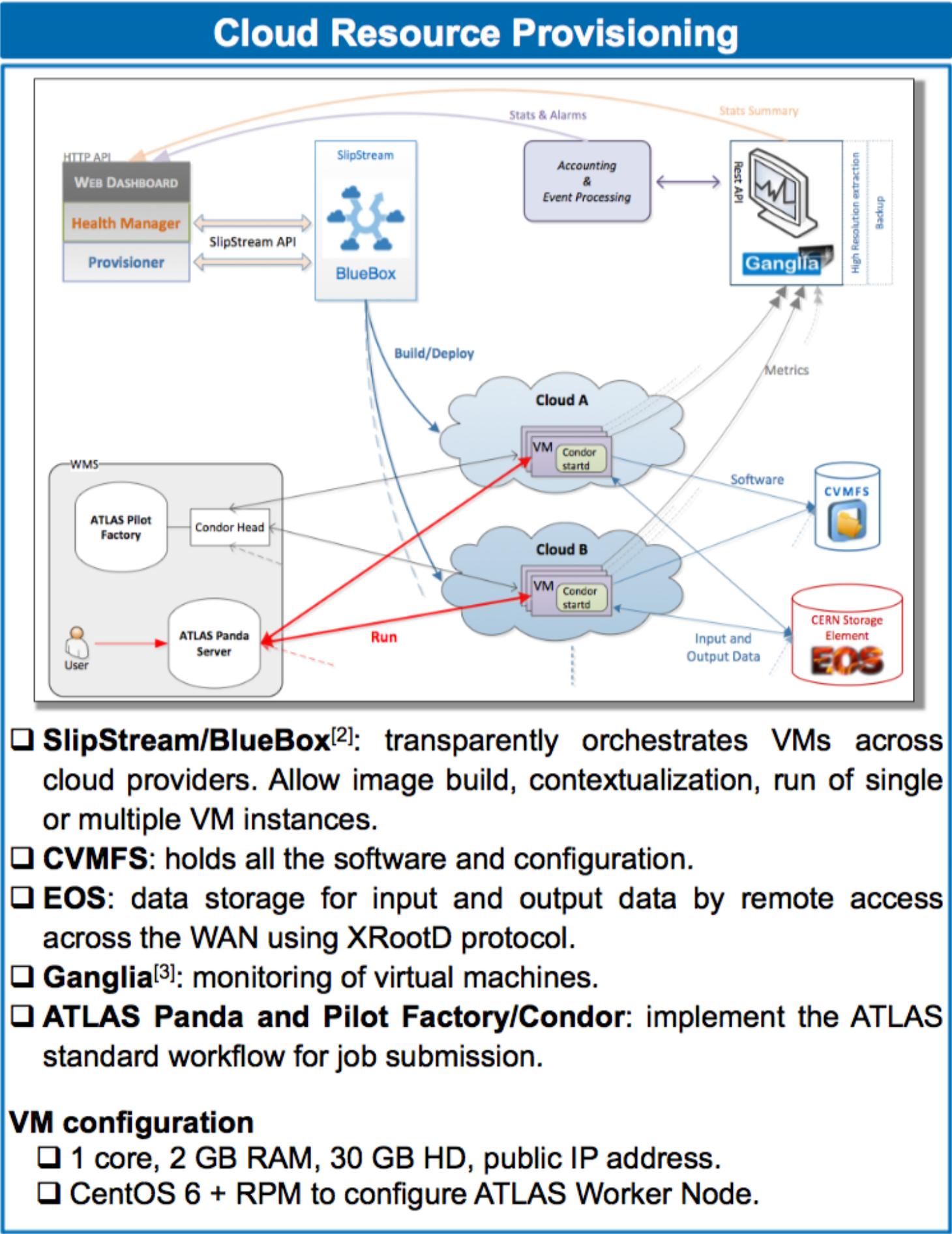
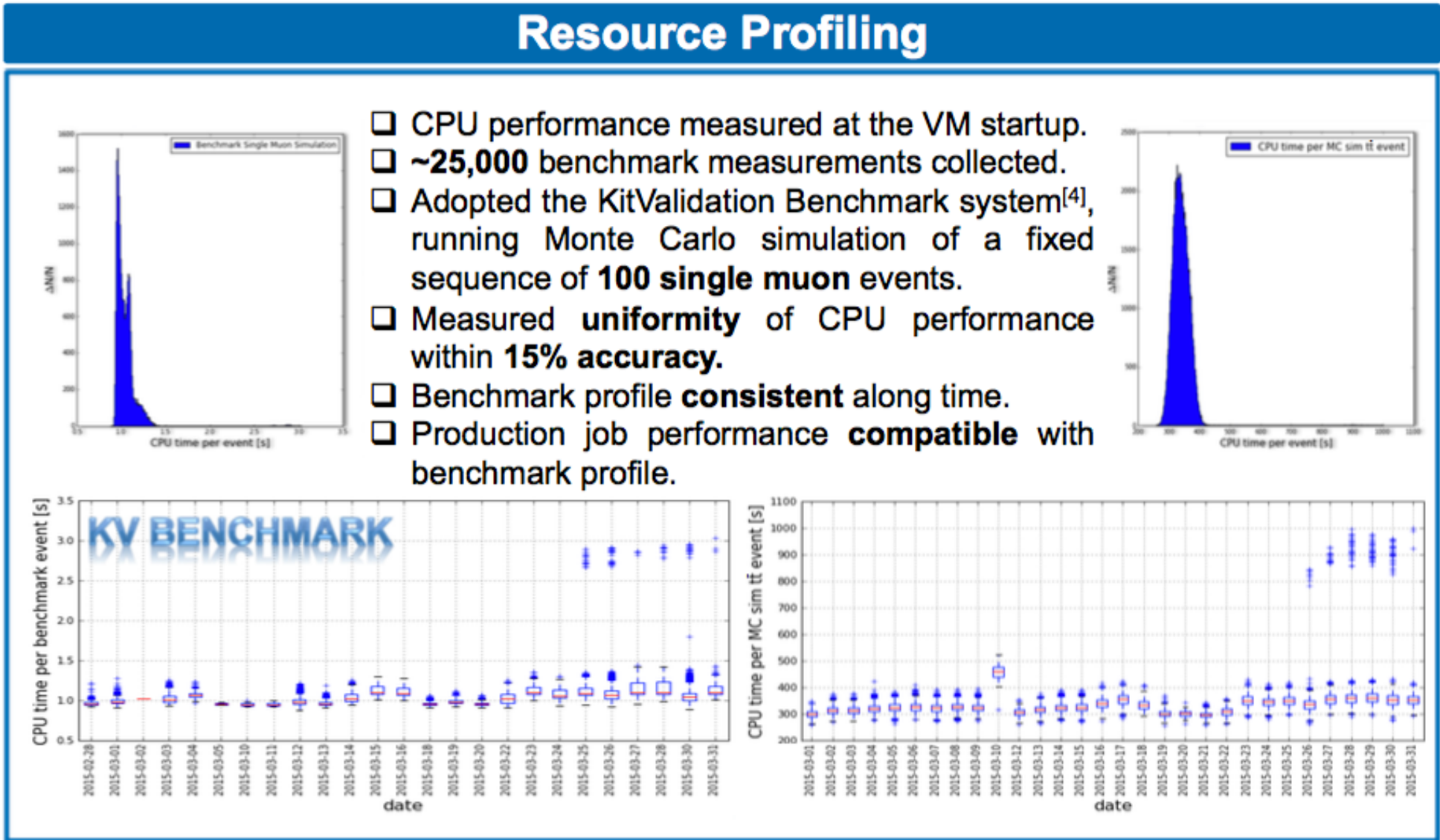
Track 7 Clouds & Virtualization - Andrew.McNab@cern.ch - CHEP 2015, 16 Apr 2015, Okinawa

6

- Ulrich presented detailed work on how to classify and benchmark VMs, and then inject performance numbers into the accounting system.
 - <https://indico.cern.ch/event/304944/session/7/contribution/86/material/slides/0.pdf>
- Andrew explained how Containers differ from Virtual Machines, and presented performance comparisons with (unoptimised) VMs, with benchmarks and with realistic HEP workloads.
 - <https://indico.cern.ch/event/304944/session/7/contribution/356/material/slides/0.pdf>
- Miguel explained how they are supporting volunteer computing with BOINC in HEP, enabled by virtualization.
 - <https://indico.cern.ch/event/304944/session/7/contribution/82/material/slides/0.pdf>
- Marek explained how identity federation is now possible between OpenStack sites
 - <https://indico.cern.ch/event/304944/session/7/contribution/226/material/slides/0.pdf>

▪ <https://indico.cern.ch/event/304944/session/15/contribution/569/material/slides/0.pdf>

Poster: Accessing commercial cloud resources within the European Helix Nebula cloud marketplace



■ <https://indico.cern.ch/event/304944/session/10/contribution/216>

Monitoring the Delivery of Virtualized Resources to the LHC Experiments

C. Cordeiro¹, A. Di Girolamo¹, L. Field¹, D. Giordano¹, D. Spiga¹, L. Villazon^{1,2}

¹CERN, ²Universidad de Oviedo

The adoption of Cloud technologies by the LHC experiments puts the burden of monitoring virtual resources upon the V.O. Monitoring the instantiated virtual machines is therefore a fundamental activity and here it is described how the Ganglia monitoring system^[1] can be exploited for monitoring and also for providing reliable information for other applications.

Cloud A, Cloud B, VMs, NAT, UDP, TCP

- Ganglia's monitoring daemon (*gmond*) sits on every Virtual Machine (VM), gathering monitoring statistics and sending them to a receiver node through unicast.
- Each provider has its own *gmond* configuration, which relies on a port number, cluster name and host (collecting *gmond*) address.
- To configure *gmond*, one can do it dynamically by fetching a pre-built JSON file, with *gmond*'s configuration from the *gmetad*'s server during the VM's contextualization.
- Communications between *gmond* are done via UDP.
- From *gmond* to *gmetad* these are done via TCP.

Deployed for ATLAS, LHCb, CMS^[2] and Helix Nebula^[3]
Monitoring >15 cloud providers
Already extracted and stored ~440 GB of raw metrics (not aggregated)

Cloud monitoring overview

Benchmarking

Ratio: [CPU Time / events]

References

- [1] Ganglia Monitoring System: <http://ganglia.sourceforge.net/>
- [2] ATLAS, LHCb, and CMS Ganglia monitors: <http://open.cern.ch/>
- [3] Accessing Commercial Cloud Resources within the European Helix Nebula Cloud Marketplace: CHEP2015 Poster 216
- [4] Cloud Accounting portal: <http://cloud-acc-dev.cern.ch/>

Contact: ganglia-devel@lists.cern.ch

MANAGING COMPETING ELASTIC GRID AND CLOUD SCIENTIFIC COMPUTING APPLICATIONS USING OPENNEBULA

The INFN-Torino Computer Centre

- Born as a WLCG Tier-2 site for the ALICE experiment at the LHC
- Then become a Tier-2 site for the BES-III experiment at IHEP, Beijing
- Now a fully virtualized cloud infrastructure comprising ~76 hosts in two clusters managed by the OpenNebula cloud controller
- Currently providing computing power to a number of applications:
 - WLCG Tier-2 sites (LHC VOs, LHCb, PANDA and others)
 - BES-III Tier-2 site (a separate middleware instance)
 - Interactive Virtual Analysis Facility for ALICE
 - Theoretical computation batch farm
 - On-demand remote medical image processing
 - Several smaller application-specific "Virtual Farms"

OpenNebula Dashboard

Elastic applications: Virtual Farms

- Usage changes with time (e.g. in bursts)
- Easy to locate idle nodes to undeploy

Anelastic applications: Grid Farms

- Work in saturated regime
- Nodes are never idle

Job duration distribution show no clear pattern

No easy way to choose multi-core VMs to undeploy

TWO PATHS TO ELASTICITY: ELASTIQ AND ONEFLOW

Elastiq is a custom Python daemon (<https://github.com/thermann/elastiq>)

- uses the EC2 interface to communicate with the cloud-controller (can work on any cloud)
- plugin implemented for HTCondor LRMS (cloud-aware)
- SCALE UP: when jobs in queue
- SCALE DOWN: when specific VM is idle

Example use-case: The ALICE Virtual Analysis Facility (VAF) (J. Phys.: Conf. Ser. 369 (2012) 012019)

- the tenant deploys 1 single VM (the master)
- Elastic configuration and workers configuration specified in master context

OneFlow is an OpenNebula tool to deploy clusters of VMs with dependencies

- designed for load balancing applications (user cannot currently decide which VMs to undeploy)
- SCALE UP: 1 VM at a time when there are queued jobs
- SCALE DOWN: when all jobs are finished

Example use-case: BESIII GRID Tier2

- master service is a CRAB CE
- slaves are DIRAC GRID worker-nodes
- LRMS is PBS (not cloud-aware)
- worker nodes publish the number of queued/running jobs to OneGate

Pros of the OneFlow approach:

- easy to configure a cluster as a single service from the OpenNebula GUI
- scale up/down manually
- change worker-node context on the fly

OUTLOOK

VM Management tools

- OneFlow in its current implementation is not optimal for this use case
- Most LRMSs used in grid sites (e.g. PBS/Torque) are not cloud-aware and cannot easily cope with nodes appearing and disappearing
- HTCondor is a better candidate
- OneFlow for large saturated use cases, Elastiq for smaller virtual farms

Scale down policies

- Large 6-8 core Virtual Worker Nodes are not ideal for this use case
- No hint from job statistics means wasted resources while the node waits for longer jobs to finish
- Need to keep some (small) WNs in draining mode all the time

Next steps

- Split the ALICE farm: static large WNs to keep the number of VMs low, smaller WNs for the elastic component
- Deploy a separate HTCondor CE for the elastic component
- Define policies and parameters for scale up and scale down of this application

The present work is supported by the Istituto Nazionale di Fisica Nucleare (INFN) of Italy and is partially funded under contract 2010P74XTM of Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale (PRIN, Italy).

Stefano Bagnasco¹, Dario Berzano², Stefano Lusso¹, Massimo Masera^{1,3}, Sara Vallero^{1,3} on behalf of the STOA-LHC project

¹ Istituto Nazionale di Fisica Nucleare; ² CERN; ³ Department of Physics, University of Torino

The DII-HEP OpenStack based CMS Data Analysis for Secure Cloud Resources

L. Osmani¹, S. Toor², M. Komu³, M. Kortelainen², T. Lindén², J. White², R. Khan⁴, P. Eerola², S. Tarkoma¹

¹University of Helsinki, ²Helsinki Institute of Physics, ³Ericsson Research, ⁴University of Alabama

Goals

Harness Grid and Cloud technologies to ensure a steady and seamless transition towards new ways of operating.

Motivation

- The High Energy Physics community is interested in performing simulations and data analysis on public or private cloud facilities.
- Currently the simulations and analysis are being performed mostly on computing and data Grids.
- The software and experience of operating on a Grid needs to be adapted for running on cloud facilities.

Contribution: Deploy a Cloud-based and Grid-enabled cluster

- We combine the elements of Cloud and Grid software components.
- We manage the VMs dynamically in an elastic fashion.
- We use the EMI authorization service (Argus) and the Execution environment Service (Argus-EES).
- Plugin developed for Argus-EES that can communicate with multiple OpenStack deployments to expand and shrink resources on-demand.
- Leverage HIP protocol for traffic management and security.

Evaluating our implementation

- The constructed virtual cluster is Cloud-based and Grid-enabled.
- OpenStack used for the Cloud and the Advance Resource Connector (ARC) for the Grid.
- Analysis software and libraries provided through CERNVMFS.
- Cloudbursting towards other community Clouds.
- CPU intensive CMS jobs CRAB jobs were run with and without the HIP protocol to study the difference in CE and worknode CPU and network usage. The jobs were run for about 170 min and each test was running for about a week on 200 concurrent jobs.

Future work

- The virtual cluster has been migrated from the test setup at the University of Helsinki to the ePouta IaaS at the CSC Kajaani Datacenter.
- The new setup is being taken into test usage.
- The largest Finnish CMS/ALICE physical cluster is planned to be replaced by this cloud setup.

Figure 1. Architecture of DII-HEP Cloud

Figure 2. DII-HEP Cloud

Figure 3. Cloudbursting

REFERENCES:

[1] White, S. Toor, P. Eerola, T. Lindén, O. Kraemer, L. Osmani, S. Tarkoma, Dynamic Provisioning of Resources in a Hybrid Infrastructure, PoS(ICSC2014)019, International Symposium on Grids and Cloud (ICSC) 2014, Taiwan.

[2] Osmani, S. Tarkoma, P. Eerola, M. Komu, M. Kortelainen, O. Kraemer, T. Lindén, S. Toor, White, An overview of the DII-HEP OpenStack based CMS data analysis. Submitted to J. Phys.: Conf. Ser. ACAST 2014.

Monitoring the Delivery of Virtualized Resources to the LHC Experiments

◉ <https://indico.cern.ch/event/304944/session/9/contribution/111>

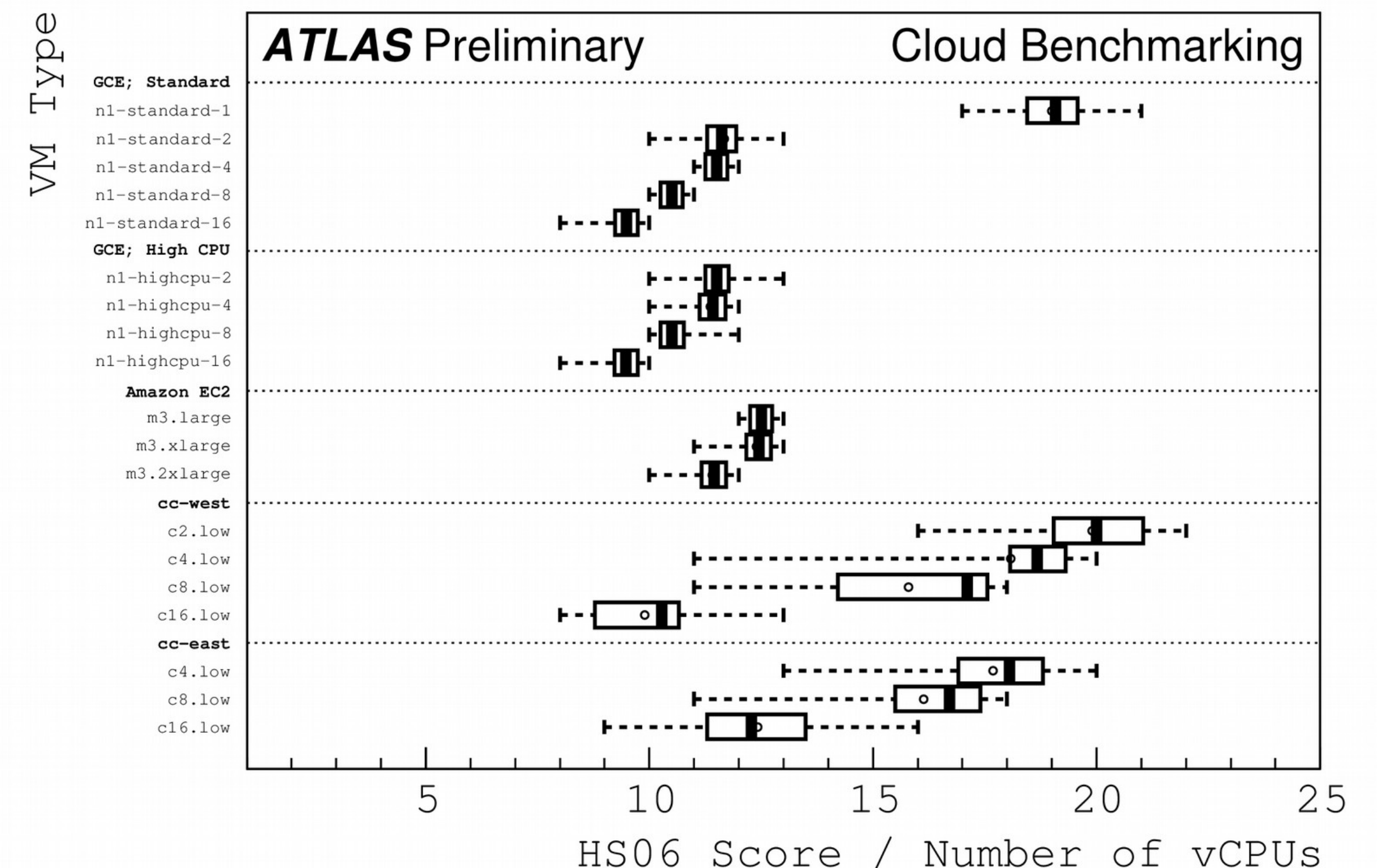
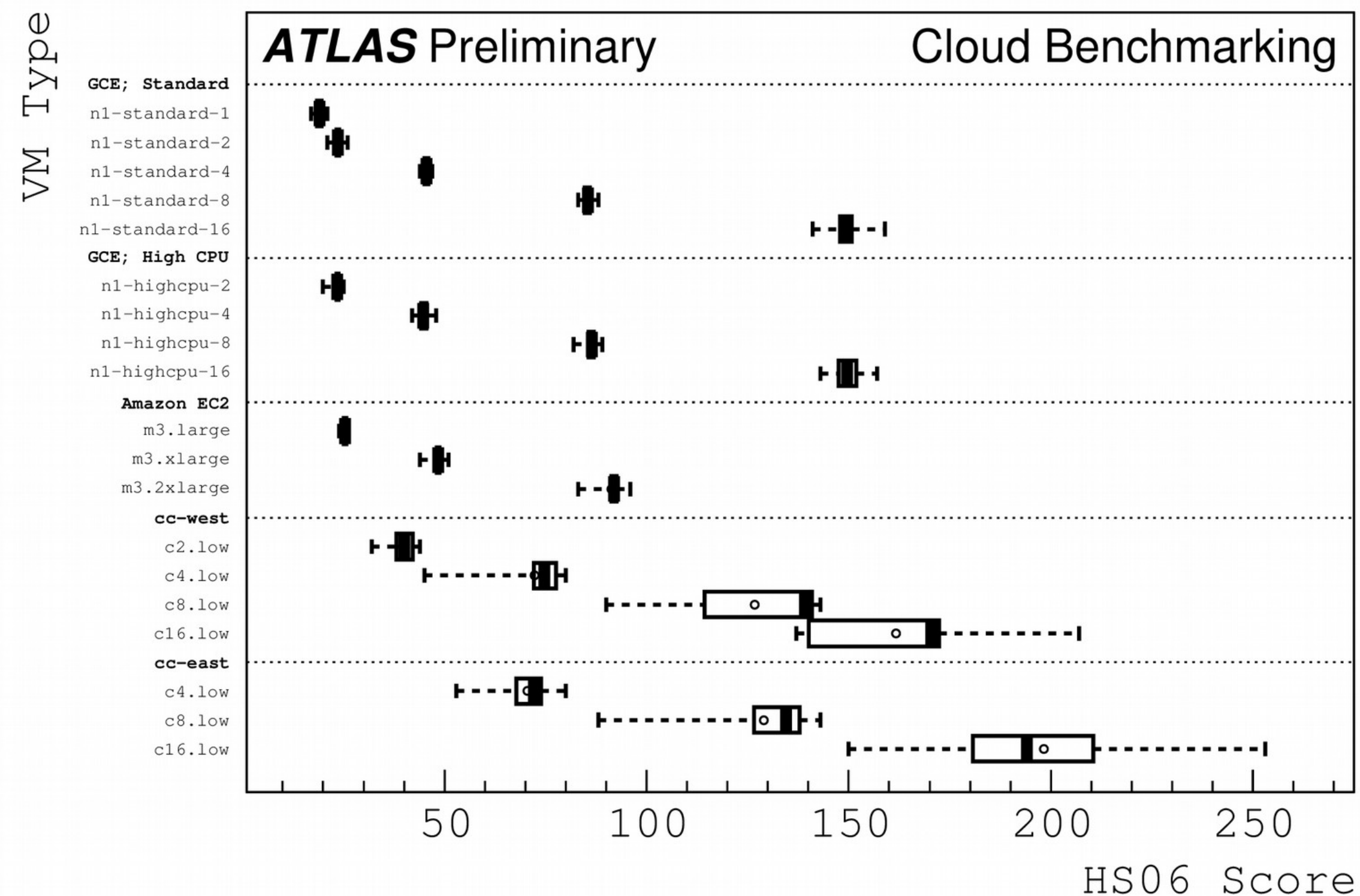
Managing competing elastic Grid and Cloud scientific computing applications using OpenNebula

◉ <https://indico.cern.ch/event/304944/session/9/contribution/387>

The DII-HEP OpenStack based CMS Data Analysis for secure cloud resources

◉ <https://indico.cern.ch/event/304944/session/10/contribution/278>

Atlas benchmarking



- General a good overview of cloud computing in Atlas, worth a read
- <https://indico.cern.ch/event/304944/session/7/contribution/146>

Cloud federation



Basic definitions

OpenStack:

- An Open Source Cloud Managing System which allows implementors to:
- Provision and manage compute, network, and storage resources quickly
 - Monitor and alert on those resources
 - Auto-scale cloud resources
 - Standardize and control disk & server images

Keystone:

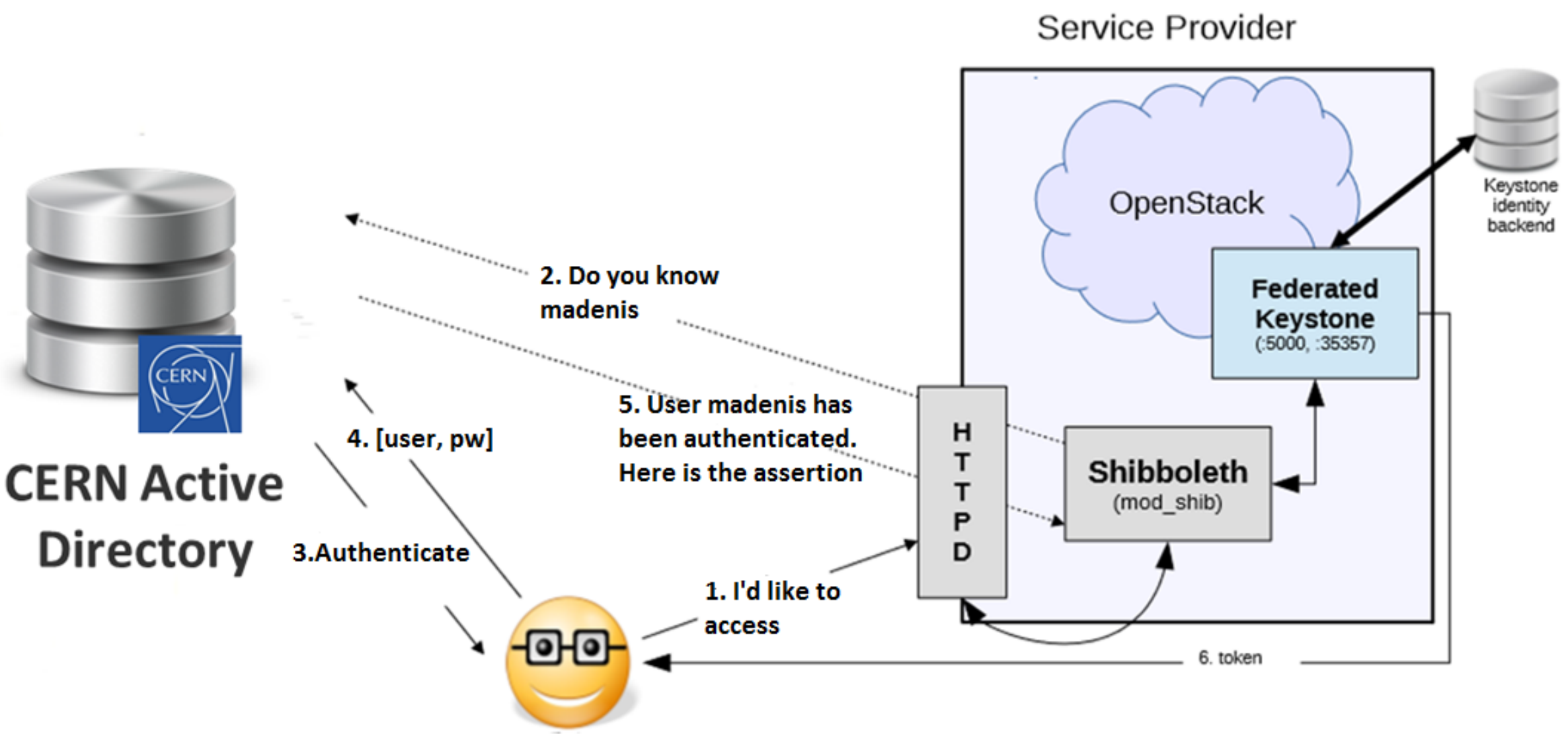
- The Identity service that comes bundled with OpenStack. Keystone allows implementors to:
- Provision users, projects, roles
 - Manage their authorization (and authentication)
 - Programmatically discover implemented cloud services

Cloud Federation:

Deployment and management of multiple external and internal cloud computing services to match business needs. A federation is the union of several smaller parts that perform a common action.



Federated authN & authZ



Credits Luca Tartarini

16/04/2015

Marek Denis – CERN openlab

7

16/04/2015

Marek Denis– CERN openlab

2

Background image: Shutterstock

Background image: Shutterstock

▪ <https://indico.cern.ch/event/304944/session/7/contribution/226>



Experiment overviews: cloud activities

- There are good overviews at CHEP for the experiments (all talks at CHEP: <https://indico.cern.ch/event/304944/timetable/?ttLyt=room#all.detailed>):
 - ◉ Atlas: <https://indico.cern.ch/event/304944/session/7/contribution/146/material/slides/0.pdf>
 - ◉ CMS: <https://indico.cern.ch/event/304944/session/7/contribution/230/material/slides/1.pdf>
 - ◉ Belle II: <https://indico.cern.ch/event/304944/session/7/contribution/294/material/slides/0.pdf>
 - ◉ LHCb: <https://indico.cern.ch/event/304944/session/7/contribution/269/material/slides/0.pdf>
 - ◉ BES III: <https://indico.cern.ch/event/304944/session/7/contribution/212/material/slides/1.pdf>

Summary Track 8: Performance increase and optimization exploiting hardware features

The Role of Memory

Performance benchmark of LHCb code on state of the art x86 architectures, R. Schwemmer et al.

We have parallel HEP data processing

→ Study Numa effects

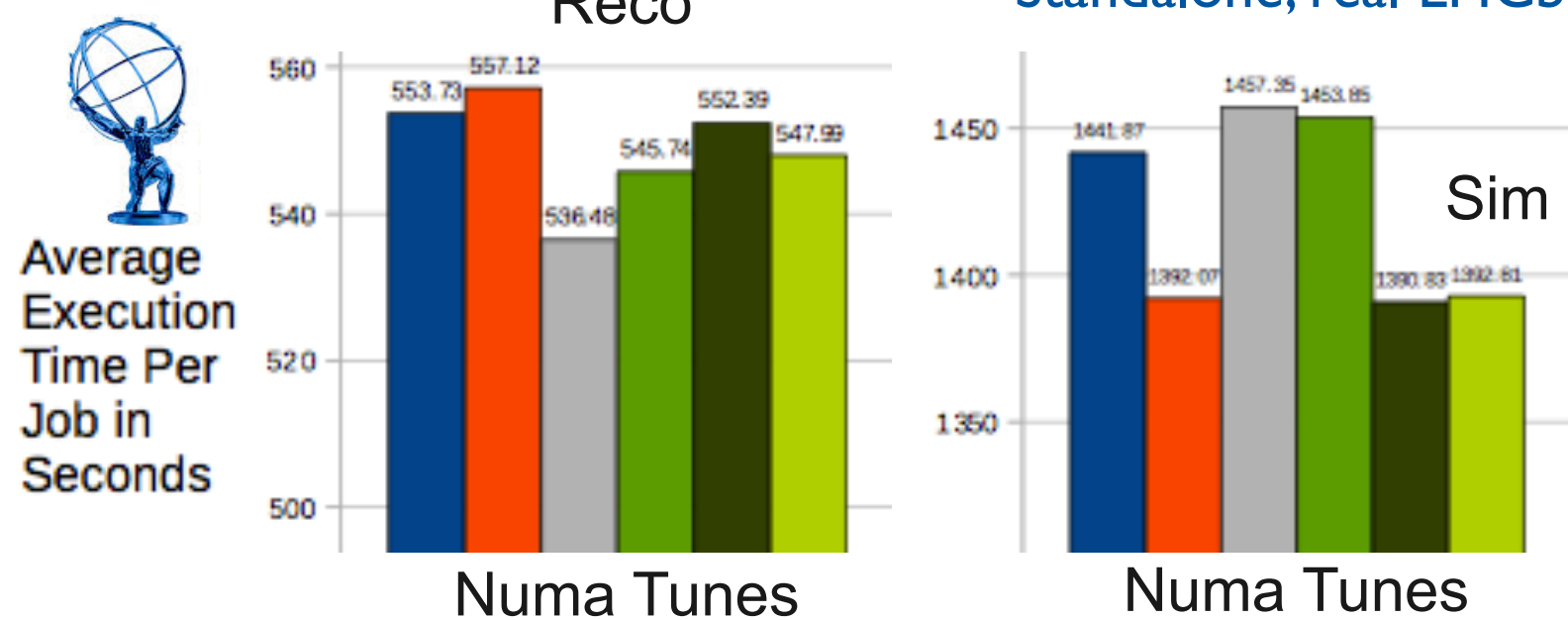
Here: no MT but MP

Spawn children on same socket: no uncore events → +14% decision rate

CPU	Decisions/s No NUMA	Decisions/s NUMA	NUMA Gain
Intel X5650 (8 cores)	599.6	648.8	1.08
Opteron X272	632.35	682	1.08
E_2630 v3 (8 cores)	865	986	1.14
E_2650 v3 (10 cores)	1129	1210	1.07



Standalone, real LHCb HLT application available on DVD!



The same tune has different effects on different workflows!

The Effect of NUMA Tunings on CPU Performance, C. Hollowell et al.

14 N. Neufeld, T. Boccali, A. Singh, D. Piparo | CHEP15 Track8 Summary

16 April 2015

■ <https://indico.cern.ch/event/304944/session/15/contribution/570/material/slides/0.pdf>

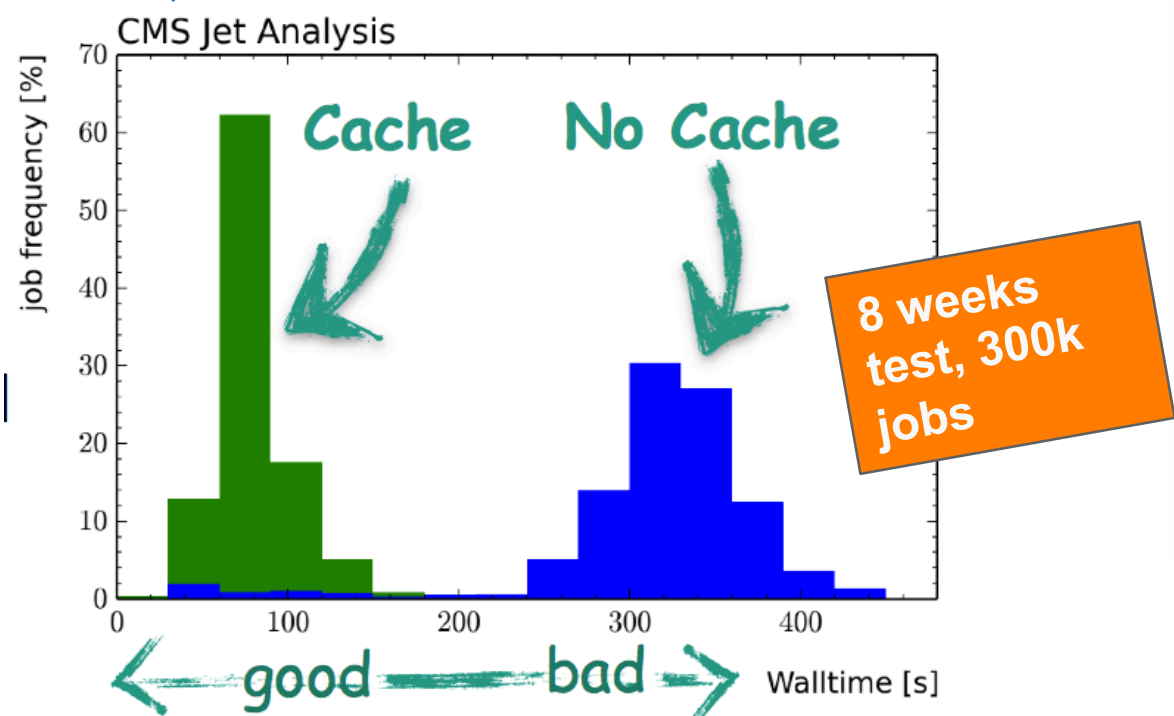
1) Strong Scaling

- FPGAs and Accelerators
- CPUs
- IO

CHEP2015 Motto:
"Evolution of Software and Computing for Experiments"

2) Throughput and HPC

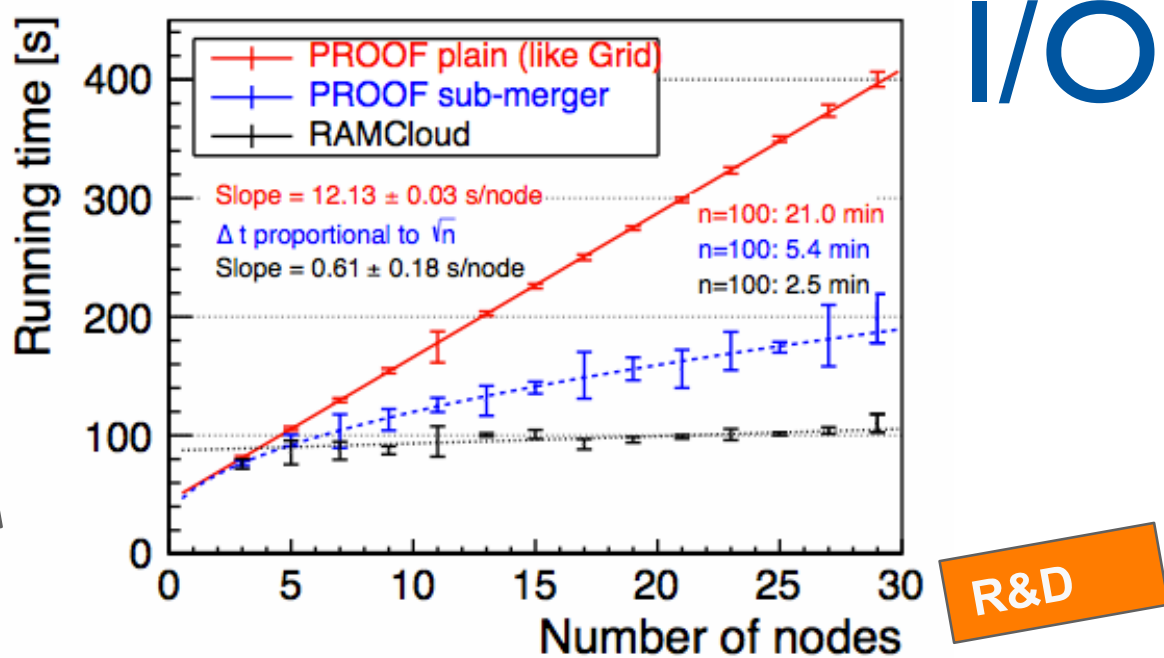
High Performance Data Analysis via Coordinated Caches, M. Fisher



University Cluster: I/O intensive HEP data analysis

Provide an SSD cache to nodes via UFS

No ad-hoc protocol, just declare consumed datasets in jdl



Histogram merging: typical serial problem

12k histos, 19M bins

RAMCloud: BigData technology

General purpose distributed storage solution competitive with highly tuned HEP tool

Large-Scale Merging of Histograms using Distributed In-Memory Computing, J. Blomer

19 N. Neufeld, T. Boccali, A. Singh, D. Piparo | CHEP15 Track8 Summary

16 April 2015

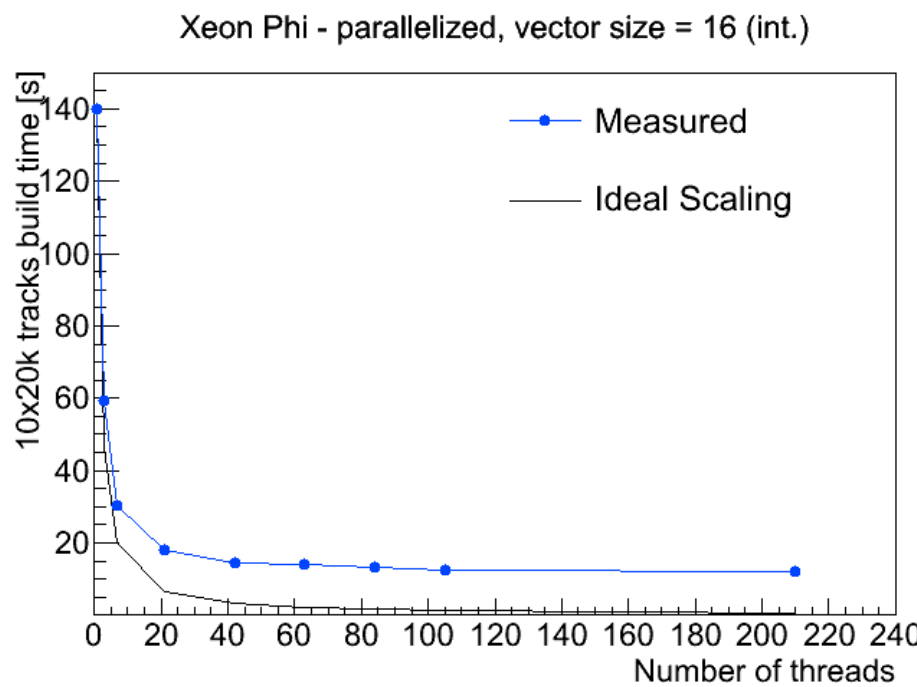
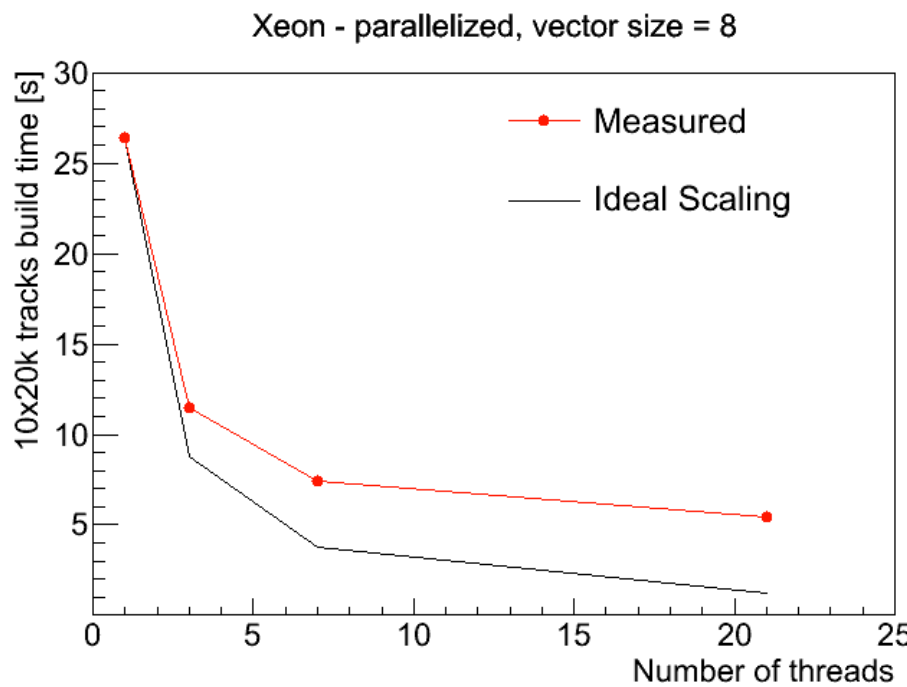


Summary Track 2: Offline Software

Kalman Filter Tracking on Parallel Architectures

#1 Mon: Reconstruction

G. Cerati



Kalman Filter based track following

- ➔ Run **full track building** with combinatorial expansion of candidates
 - ▶ ultimate physics performance, slower
 - ▶ 85% (95%) of tracks found with $\geq 90\%$ (60%) of the hits
- Parallelization is implemented by **distributing threads across 21 eta bins**
 - ▶ for nEtaBin multiple of nThreads, split eta bins in threads
 - ▶ for nThreads multiple of nEtaBin, split seeds in bin across nThreads/nEtaBin threads
- Large **speedup** achieved, both on Xeon and Xeon Phi
 - ➔ up to **~5x on Xeon and >10x Xeon Phi**
 - ➔ speedup saturates above nThreads=42

Summary:

- Significant speedup achieved both on Xeon and Xeon Phi.
- Ideal scaling indicates a large margin for further improvements.

R is an open source language and environment for statistical computing and graphics. ROOT-R package gives access to ROOT users to the R capabilities and its rich functionality.

Using R in ROOT with the ROOT-R Package

#6 Thu: Tools

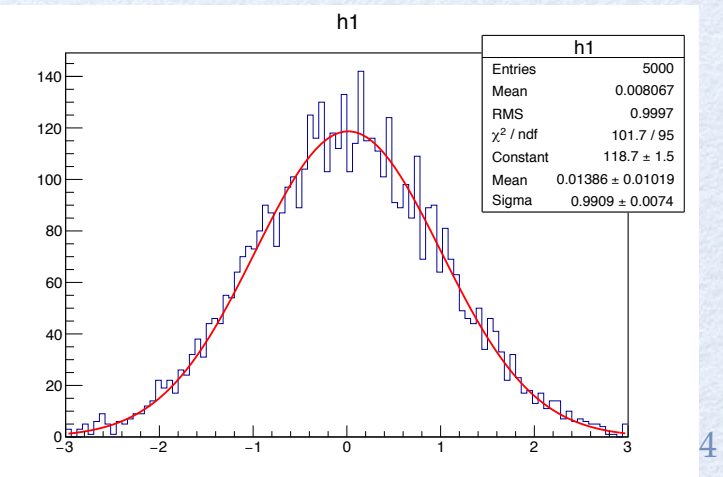
L. Moneta

Example of RMinimizer

- ROOT plugin for Minimisation implemented using R
 - *developed by Kirby Hermann (GSOC student 2014)*
 - give access to R optimisation tools when fitting or multi-dimensional function minimisation
 - based on R optim and optimx packages

```
ROOT::Math::MinimizerOptions::SetDefaultMinimizer("RMinimizer", "L-BFGS-B");  
hist->Fit("gaus");
```

```
root [4] h1.Fit("gaus")  
Value at minimum =101.673  
*****  
Minimizer is RMinimizer / L-BFGS-B  
Chi2          =      101.673  
Ndf           =         95  
NCalls        =        265  
Constant      =    118.694 +/- 1.47659  
Mean          =    0.0138555 +/- 0.0101907  
Sigma         =    0.990906 +/- 0.00741443
```



Summary:

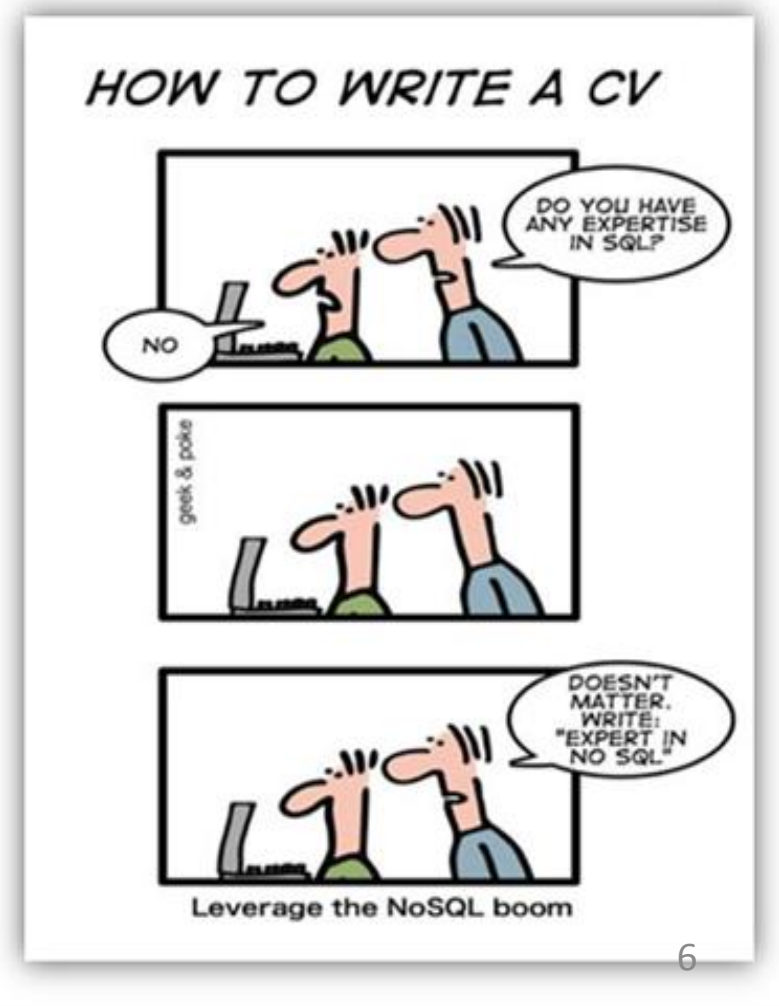
- ROOTR provides easy access to R tools in ROOT and C++
- Easy to use directly R from ROOT prompt
- Package is ready to be released in the next ROOT production version (6.0.4)

▪ <https://indico.cern.ch/event/304944/session/15/contribution/572/material/slides/0.pdf>

Summary Track 3: Databases, Storage, ...

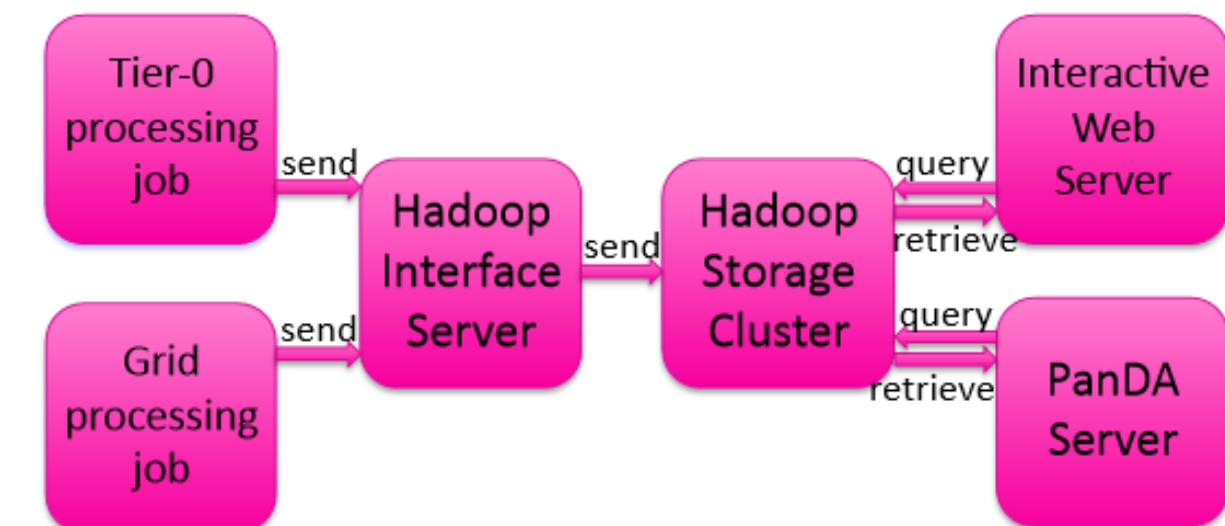
DB - New technologies evaluation

- NoSQL technologies for the CMS Conditions Database (R. Sipos)
- Evaluation of NoSQL databases for DIRAC monitoring and beyond (F. Stagni)
- Studies of Big Data meta-data segmentation between relational and non-relational databases (M. Golosova)
 - Performance evaluation of various products: InfluxDB, OpenTSDB, ElasticSearch, MongoDB, Cassandra, RIAK
 - Goal - optimize search, increase speed, aggregate (meta)data, time series
 - All evaluations show promising and better results, compared to standard SQL technologies
 - Some are about to be put in production
 - More work ahead...



Event Indexation and storage

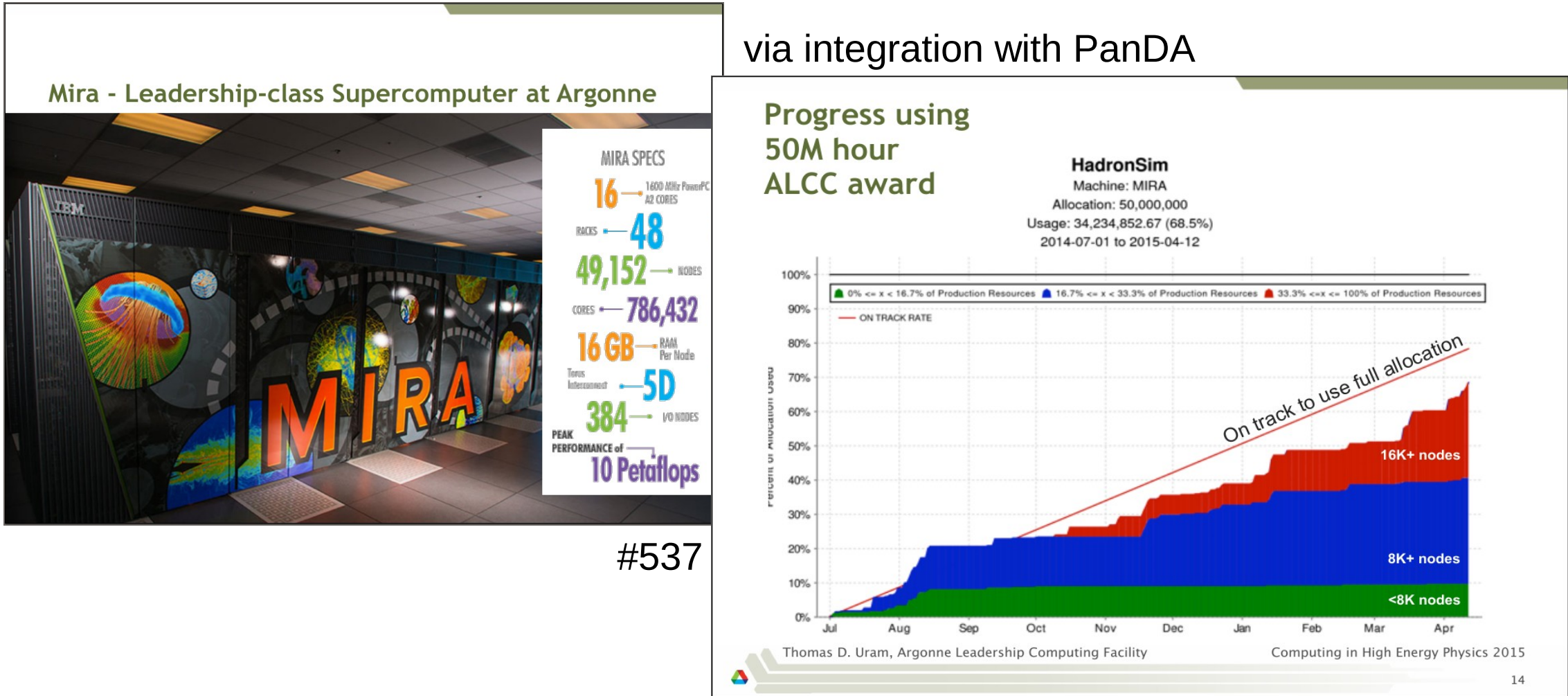
- The ATLAS EventIndex: architecture, design choices, deployment and first operation experience (D. Barberis)
- Distributed Data Collection for the ATLAS EventIndex (J. Sanchez)
 - Complete catalogue of *all* events in ATLAS
 - Billions of events amounting to 2TB of raw information for Run 1 (twice as much for Run20)
 - Web Service search - seconds for a search on a key, minutes for complex queries
 - Uses Hadoop at CERN
 - ActiveMQ as messaging protocol to transport info between producers and consumer, data encoded in JSON



▪ <https://indico.cern.ch/event/304944/session/15/contribution/573/material/slides/1.pdf>

Middleware Job Management / Pilots

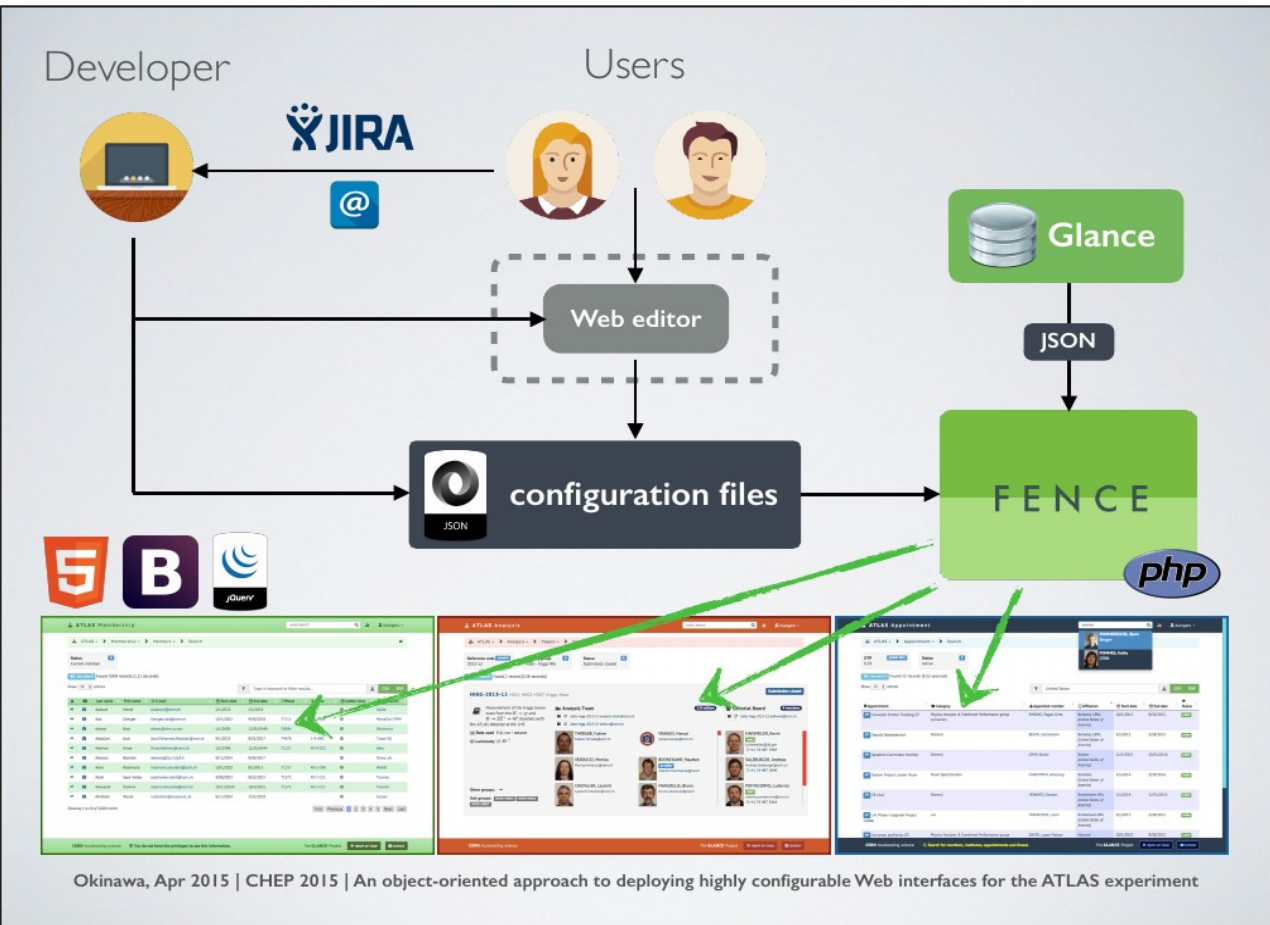
- We have seen how MIRA became the primary Alpgen event generation site for ATLAS



10

Tools

- From ATLAS we saw an interesting new way of developing web interfaces

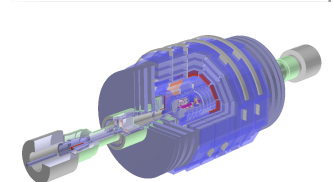
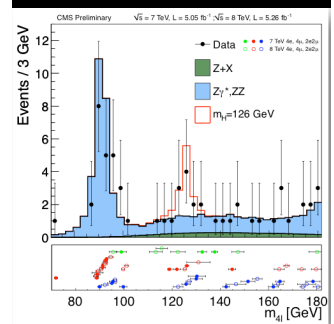
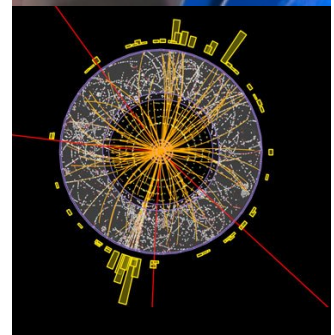
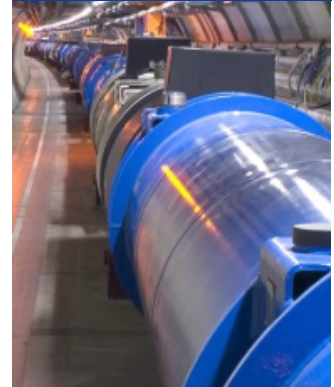
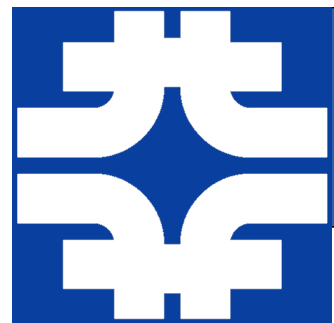


#167

36

▪ <https://indico.cern.ch/event/304944/session/15/contribution/574/material/slides/0.pdf>

Reports on ROOT 6 and beyond



Philippe CANAL
root.cern.ch

Conclusion

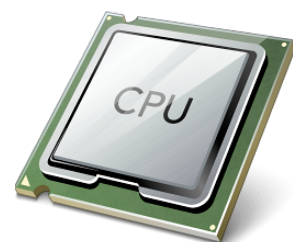
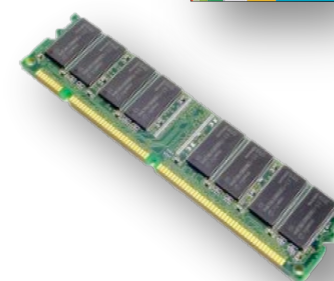
- **ROOT** Modernization underway
 - Starting to add **new** API that will overtime replace then deprecated historical API
 - Making writing [physics [analysis]] code even simpler, more intuitive and more robust
- Main Driving Principles
 - Simplicity
 - Robustness
 - Performance
 - Embrace multi-tasking and vectorization
 - Provide even better features
 - Continue our many collaborations (e.g. *Python*, *R*, *I/O*)



Where Are We Now: CMS Example

Memory

- pp→ttbar events @ 13 TeV (event loop):
 - Generation & Simulation: **-6% RSS** wrt ROOT5
 - Yes, better than ROOT5 ☺
 - Reconstruction: **+4% MB RSS** wrt ROOT5
- RSS variations: depend on **amount of interpreted functions**
 - E.g. cuts specified in job configuration



Runtime: ~Identical in the event loop

Also thanks to experiments' flexibility and willingness to make this happen – thank you!

CHEP 2015 - Okinawa

13 April 2015

19

16/4/2015

CHEP2015, Okinawa – Track 4

13

- Many updates and improvements, Impressive work on optimization and validation, Path to API transformation. Strong collaboration with experiments.
- <https://indico.cern.ch/event/304944/session/4/contribution/411/material/slides/1.pdf>

Summary Track 6: Facilities, Infrastructure, Network

Session 2: Tue 14:00 h – 16:00 h [C209] → Convener: H. Meinhard

- GridPP – preparing for Run-2 and the wider context ([Jeremy Coles](#) / Cambridge)
- Getting prepared for the LHC Run2: the PIC Tier-1 case ([Josep Flix](#) / PIC)
- Scheduling multicore workload on shared multipurpose clusters ([Jeff Templon](#) / NIKHEF)
- Active job monitoring in pilots ([Manuel GIFFELS](#) / KIT)
- Migrating to 100GE WAN Infrastructure at Fermilab ([Phil DeMar](#) / FermiLab)
- Monitoring WLCG with lambda-architecture: a new scalable data store and analytics platform for monitoring at petabyte scale ([Luca Magnoni](#) / CERN)
- A Model for Forecasting Data Centre Infrastructure Costs ([Renaud Vernet](#) / CC-IN2P3)
- High-Speed Mobile Communications in Hostile Environments ([Stefano Agosta](#) / CERN)



Trends from the track

- Consolidation to using more flexible/scalable tools, at all levels
- Costs optimizations at the Facilities & Infrastructures
 - Transition of established services (as of today) to more flexible solutions
 - Services asy-to-operate, and at scale (i.e. changes at many sites)
 - Even, a paradigm change in the way we exploit resources at the sites
- Increased networks around (full deployment of 100 GBE elsewhere)
 - Network is becoming more and more important → we'll have better use/monitors
 - IPv6 for sure... SDNs / Named Data Networks in production?
- Better tools and techniques to monitor “data”
 - subject to revision cycles: what's good today, cannot be used in the future (challenge)
- Better tools for us to better interact, share our results, share knowledge
 - which will positively impact in the quality of the work done

17th April 2015

CHEP 2015 沖縄本島 – Summary Track6 – P. DeMar, [J. Flix](#), P. Hristov, H. Meinhard, E. Yen

■ <https://indico.cern.ch/event/304944/session/15/contribution/575/material/slides/0.pdf>

- A few computing sites presented their readiness for LHC Run2
 - Required substantial R&D and tuning: new protocols, data federations, ..
- Costs concerns / **reducing costs**:
 - Detailed studies on estimated efforts for site tasks
 - Detailed studies on hardware costs, electricity costs and trends
 - Infrastructure improvements (free-cooling techniques, UPS upgrades, ...)
 - Detailed plans / costs for network upgrades
 - Deployment of more flexible, cost effective and easy ways to operate services
 - Batch system consolidation (HTcondor, moving from CREAM-CEs,...)
 - Cloud resources R&D / VAC context
 - People operate/run the services, perform R&D, tuning, ask/drive projects
 - **You(we) don't want to lose people!**

Monitoring



Big Data Analytics as a Service Infrastructure: Challenges, Desired Properties and Solutions

Manuel Martín Márquez
CERN- European Centre for Nuclear Research

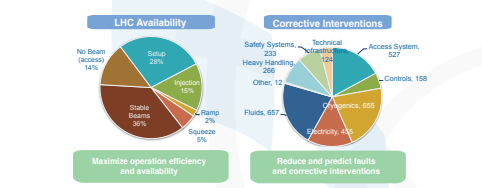
Manifesto

CERN's accelerator complex and detectors are an extreme data generator, every second an important amount of comprehensively heterogeneous data coming from control equipment and monitoring agents is persisted and needs to be analysed. Over the decades, CERN has applied different approaches, techniques and technologies. This has minimized the necessary collaboration to deliver cross data analytics over different domains. Essential to unlock hidden insights and correlations between the underlying processes, which enable better and more efficient daily-based accelerators operations and more informed decisions.

The proposed Big Data Analytics as a Service Infrastructure aims to: (1) Integrate the existing developments, (2) Centralize and standardize the complex data analytics needs for the CERN's research and engineering community, (3) Deliver real time and batch data analytics capabilities and (4) provide transparent data access and extraction-transformation-load, ETL, mechanisms to the different and mission-critical existing data repositories.

Data Analytics Objectives

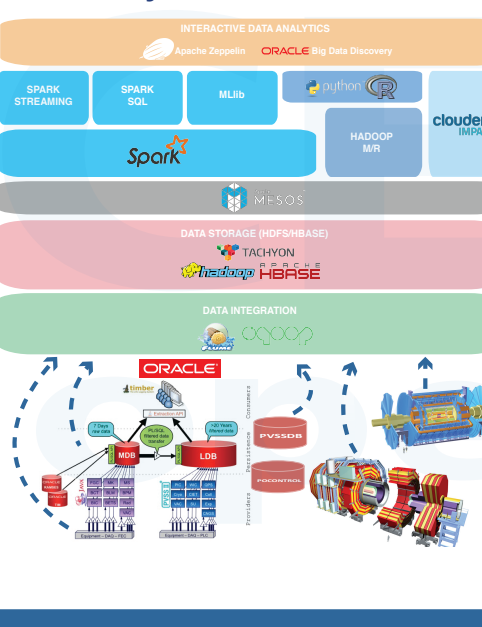
Optimize CERN's Controls:



Evolving CERN's Control and Monitoring Systems



Data Analytics as a Service



Technical Challenges

Data Access and Repositories Integration
Persist large amount of heterogeneous data
-Cryogenics, vacuum, power converters...
Millions of control devices (time series data)
-Sensors, actuators, monitoring agents
Integrate existing control data repositories
Provide transparent and flexible data access
Near-Real-Time processing
Order of GBs per second - Low latency
Integrate pre-existing knowledge and inferred
Scalable and fault-tolerance
Batch and micro-batch analysis
Integrate different tools and frameworks


Educational Aspects

Data Scientist - General
New Professional Profile
Many domains of expertise involved
Data Scientist - CERN
Need to train engineering and control teams

Some Use Cases

Faulty cryogenics valves detection
Signals used:
S = aperture order - aperture measured
Features extractions based on S
-Variance
-Percentile 99.9
-Rope distance - R(S)
-Noise Band - B(S)


Automatic faulty valves detection system
SVM - Support Vector Machine
Anomaly detection on beam screen cryogenics control
PID output (time series) segmentation
Segments characterization
Features extraction
Classification based on features

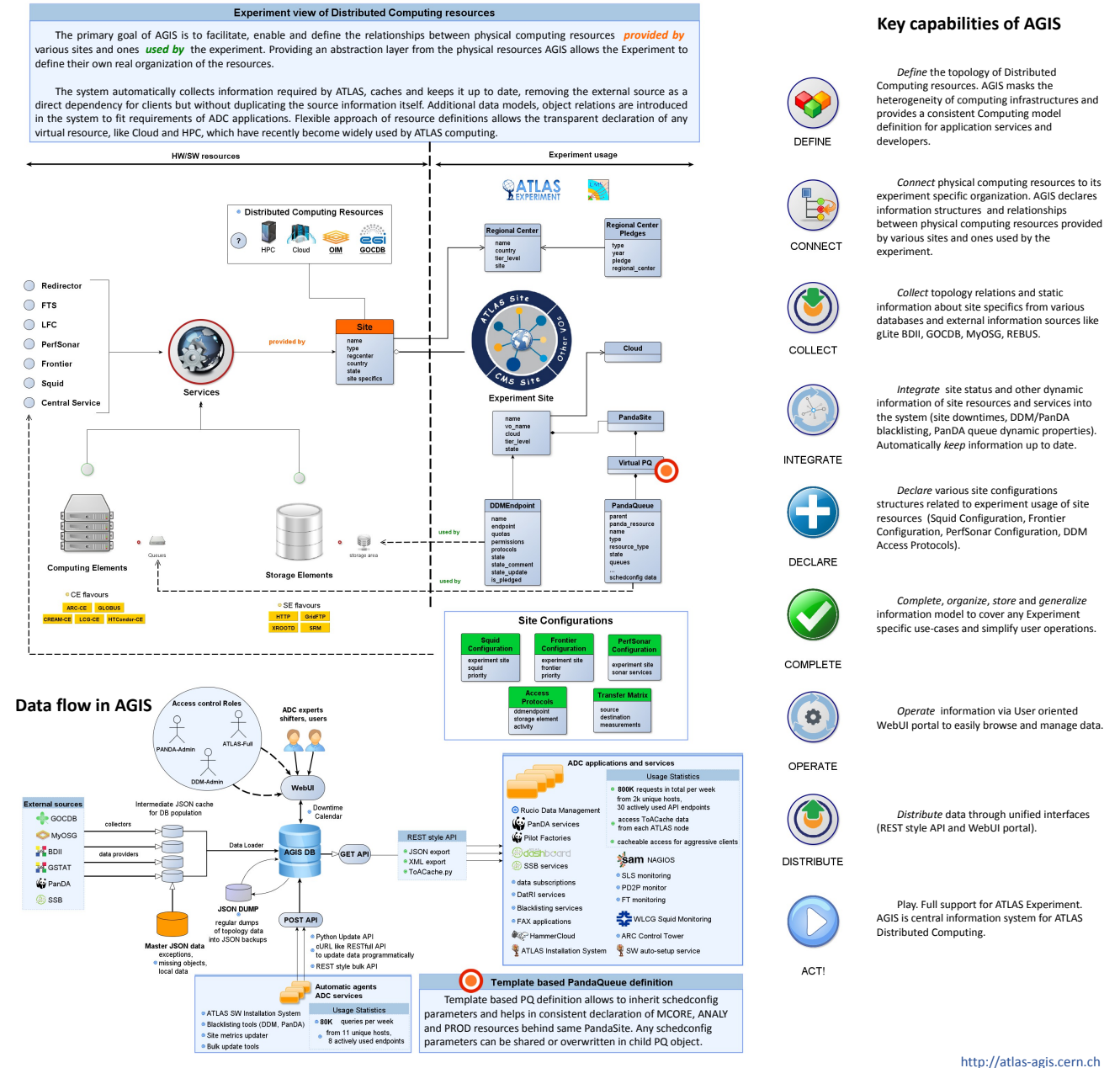


AGIS: Evolution of Distributed Computing Information system for ATLAS

A. Anisenkov¹, A. Di Girolamo² on behalf of the ATLAS Collaboration
¹ Budker Institute of Nuclear Physics, Novosibirsk, Russia
² CERN, Geneva, Switzerland

AGIS is the information system designed to integrate configuration and status information about resources, services and topology of the computing infrastructure used by ATLAS Distributed Computing (ADC) applications and services. Being in production during LHC Run 1 AGIS became the central information system for Distributed Computing in ATLAS and it is continuously evolving to fulfill new user requests, enable enhanced operations and follow the extension of ATLAS Computing model.





AGIS functionalities allow the ADC community, experts and shifters to configure and operate production ADC systems and Grid applications. AGIS is evolving towards an experiment-non-specific information system.

Poster presented at the 21st International Conference on Computing in High Energy and Nuclear Physics (CHEP), Okinawa, Japan, April 12-17, 2015.

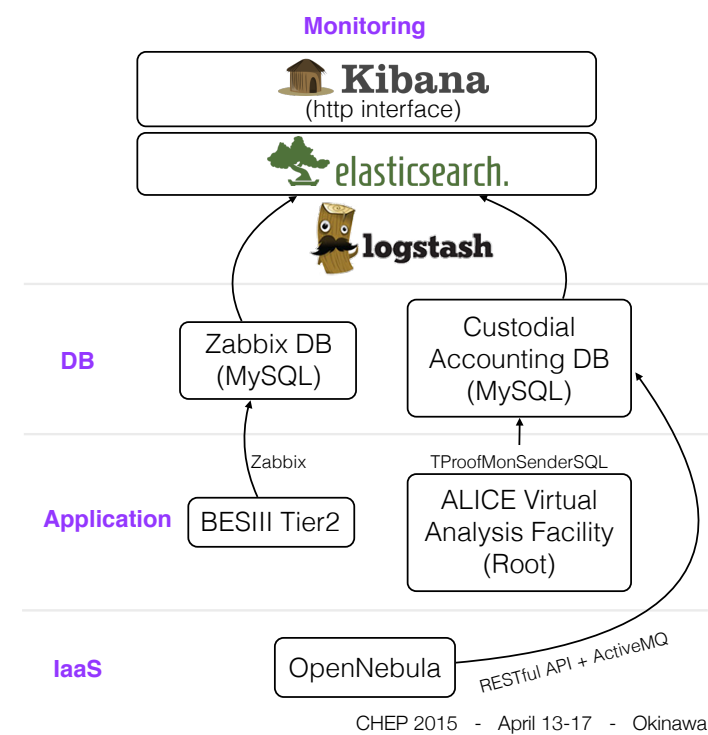
- Big Data Analytics as a Service Infrastructure/ Challenges, Desired Properties and Solutions

- <https://indico.cern.ch/event/304944/session/9/contribution/65>

- AGIS/ Evolution of Distributed Computing information system for ATLAS

- <https://indico.cern.ch/event/304944/session/10/contribution/168>

General set-up

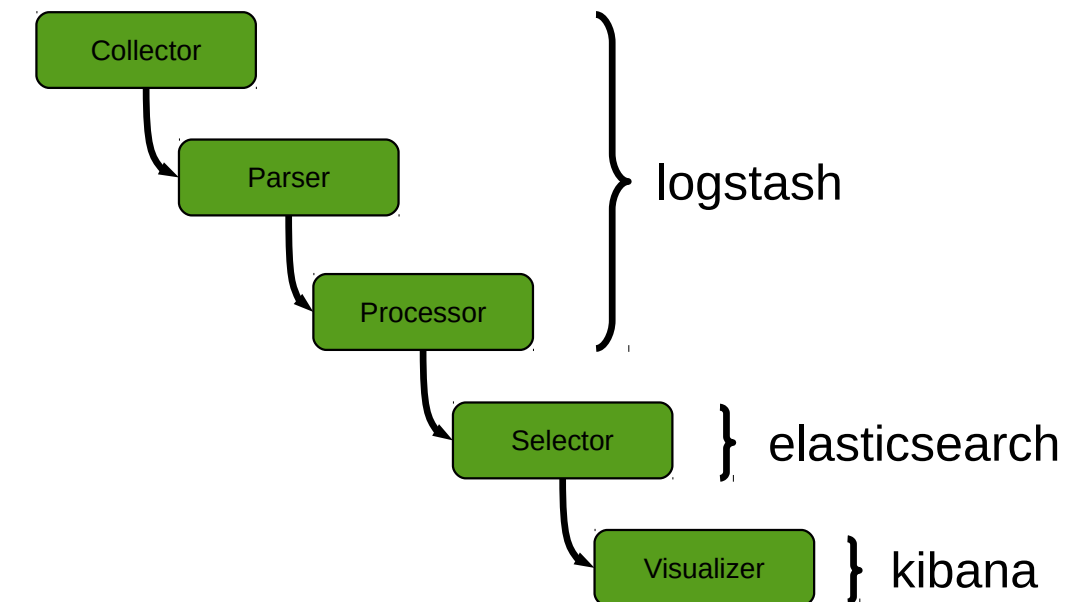


- data stored in high-availability MySQL server
- redundant step, but allows for more flexibility
- Italian Grid accounting now dismissing MySQL
- the framework was developed to monitor user activity within the VAF
- as a proof of concept also retrieve info from Zabbix DB, some custom view was created to ease indexing

- Integrated Monitoring-as-a-service for Scientific Computing Cloud applications using the Elasticsearch ecosystem

- <https://indico.cern.ch/event/304944/session/7/contribution/389>

The Flow



Tigran Mkrtchyan | Visualization of dCache accounting information | Date | Page 9



- Visualization of dCache accounting information with state-of-the-art Data Analysis Tools

- <https://indico.cern.ch/event/304944/session/4/contribution/45>

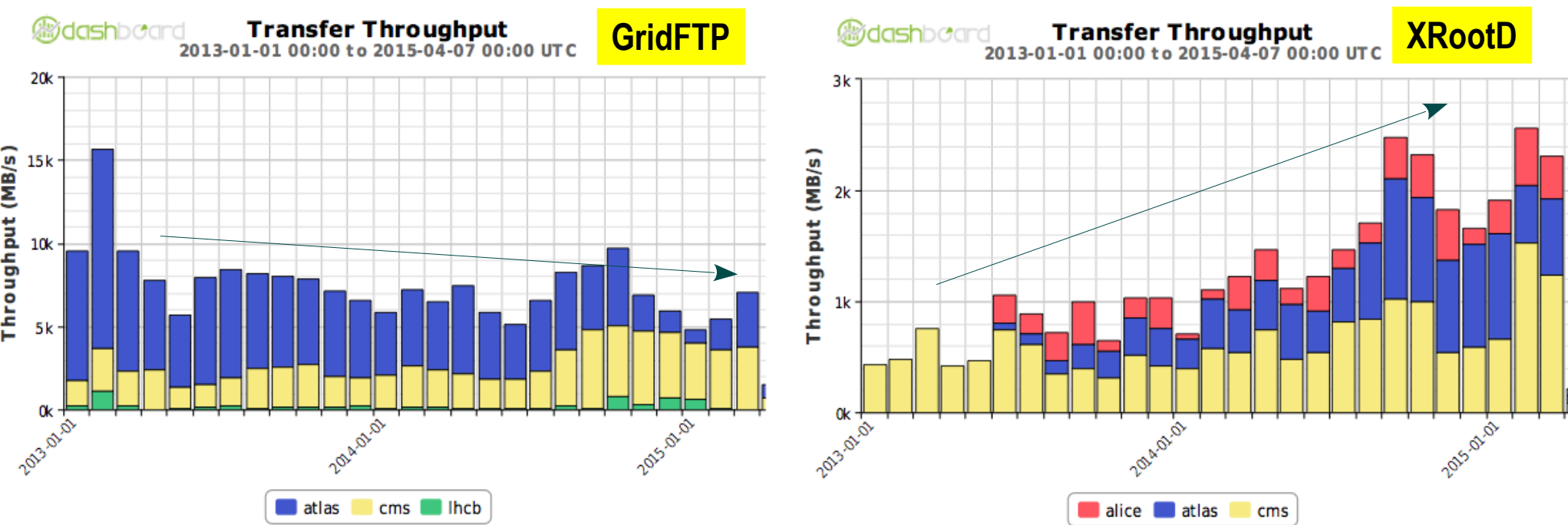
- WLCG workshop 2015 Highlights

WLCG Status and Readiness for Run2



Most relevant achievements 1/3

- Deployment at the WLCG sites of the **XRootD federated data storage** for the **FAX** (ATLAS) and **AAA** (CMS) projects
 - Monitoring, third party plugins, SAM tests, deployment instructions



XRootD traffic is increasing: accounts for 1/3 of the GridFTP traffic

11/Apr/2015

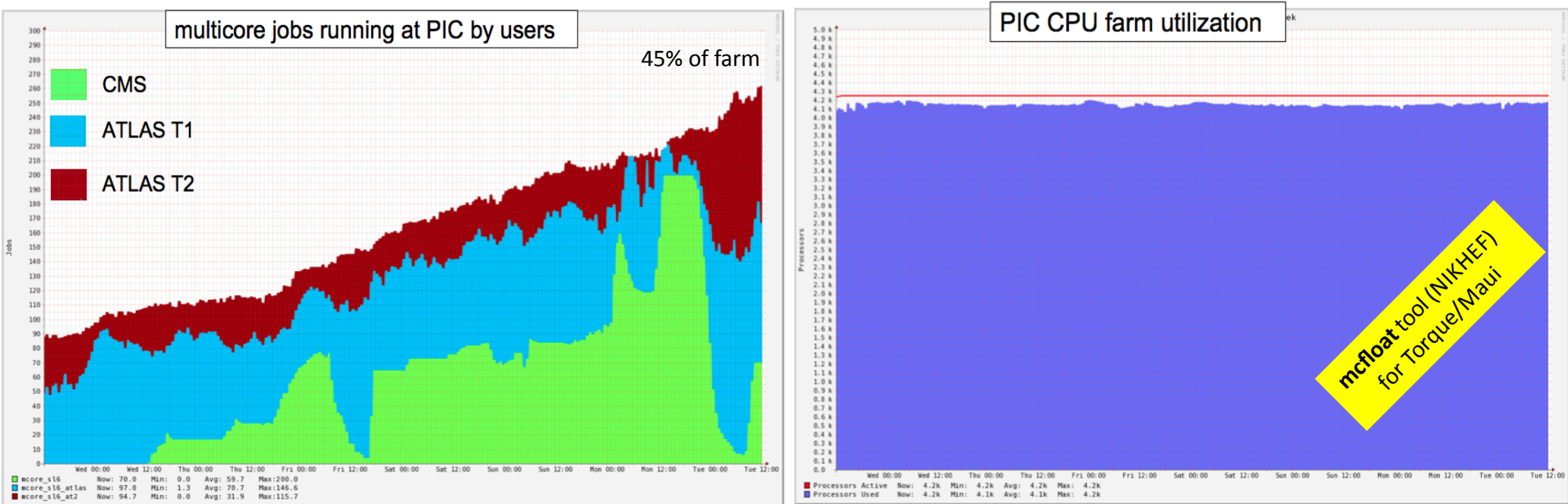
WLCG Collaboration Workshop, Okinawa, Japan

6



Most relevant achievements 3/3

- Successful shared use of the common resources in **multicore mode** by ATLAS and CMS has been achieved
 - Discussions to optimize the usage in popular batch systems (dynamic partitioning)
 - Paved the way to exploit multicore resources efficiently



Controlled ramp up of multicore resources reduces draining impact on farm utilization
98% full farm while ramping up under combined pressure

11/Apr/2015

WLCG Collaboration Workshop, Okinawa, Japan

9





▪ <http://indico.cern.ch/event/345619/session/0/contribution/7/material/slides/0.pdf>

WLCG Status and Readiness for Run2



WLCG Operations :: facing Run2

MESSAGE: WLCG Ops & the Experiments are ready for Run2

- What can we **expect** to occur during Run2?
 - **Middleware support** could become an issue, if unattended
 - Migrations to **new batch systems** at the sites 
 - * Tier-0 and some Tier-1s already started tests and drafted deployment plans
 - Migration to a **new OS**?
 - Passing **job parameters** to batch systems to optimize scheduling (ATLAS)
 - * Events complexity → More multicore & high memory jobs
 - More **demanding network & IPv6 deployment** 
 - * Better monitoring & troubleshooting / increase of *bandwidth* capacities at sites
 - * IPv6 compliance of the middleware and the experiment software is crucial to allow for IPv6-only resources
 - More **flexible ways to exploit resources** at Tier-1s/Tier-2s 
 - * And use of opportunistic resources, including HLT farms during LHC operations
 - Exploitation of **Cloud Resources** in regular Ops? (commercial/public) 

11/Apr/2015

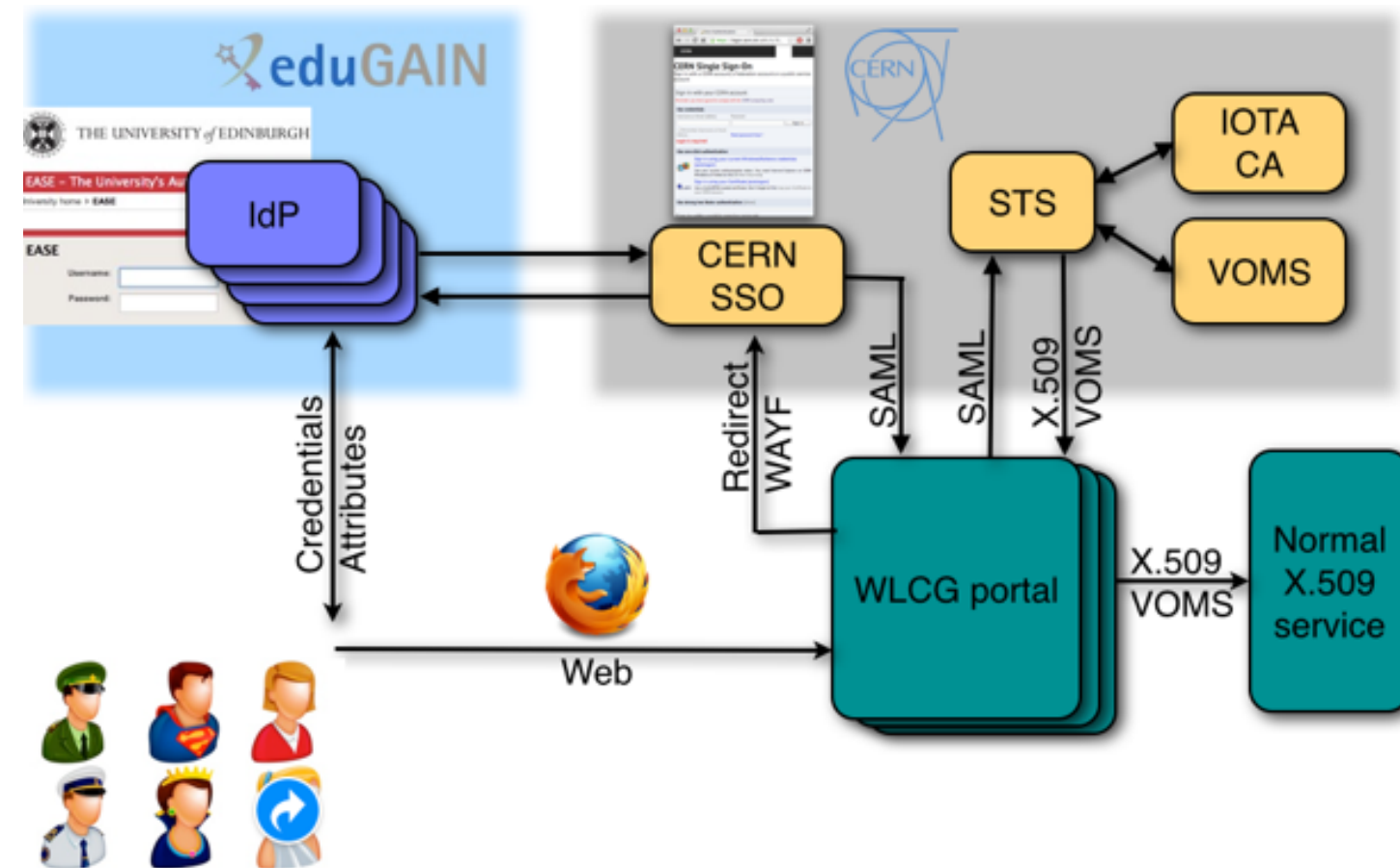
WLCG Collaboration Workshop, Okinawa, Japan

21

▪ <http://indico.cern.ch/event/345619/session/0/contribution/7/material/slides/0.pdf>

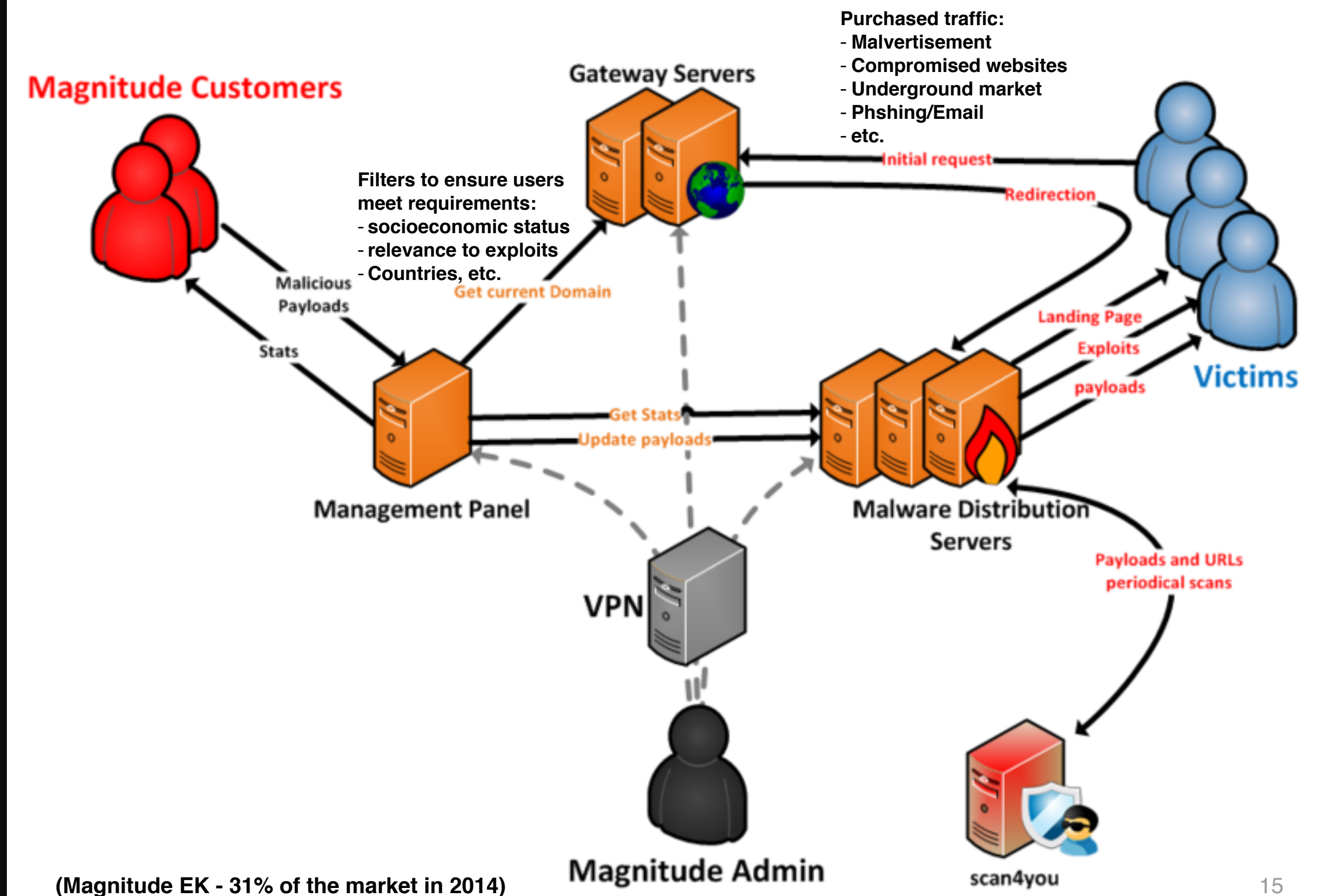
Identity federation

- Identity management basic building block
- Expected in every future computing e-infrastructure services
- No longer one account per service ; Just **one** global identity
- (now hopefully) familiar goal:



4

Malware-as-a-service



15

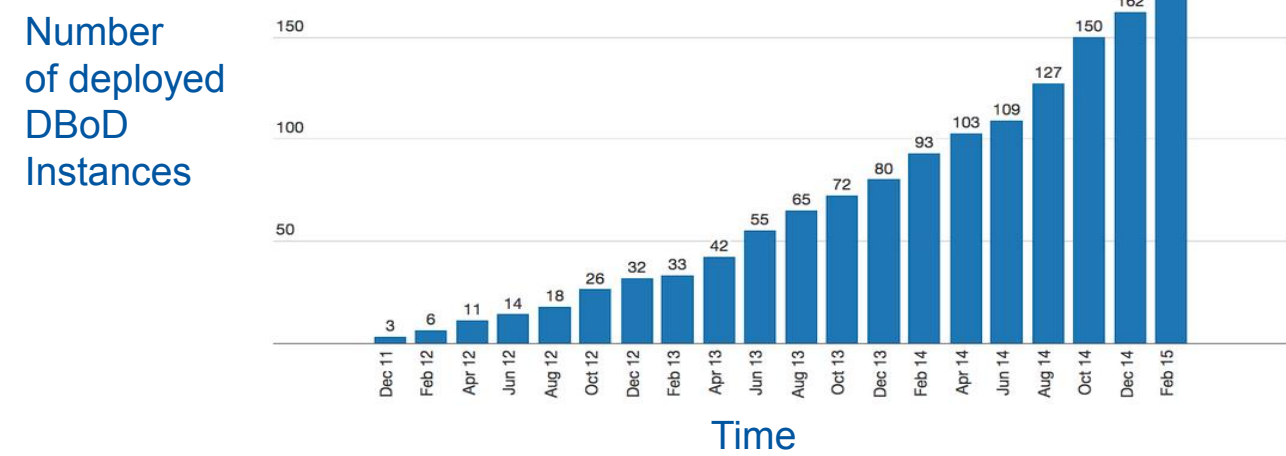
- Very good overview of what is going on in security especially in the world in connection to our scientific efforts:

◦ <http://indico.cern.ch/event/345619/session/0/contribution/0/material/slides/0.pdf>

Database services during Run2

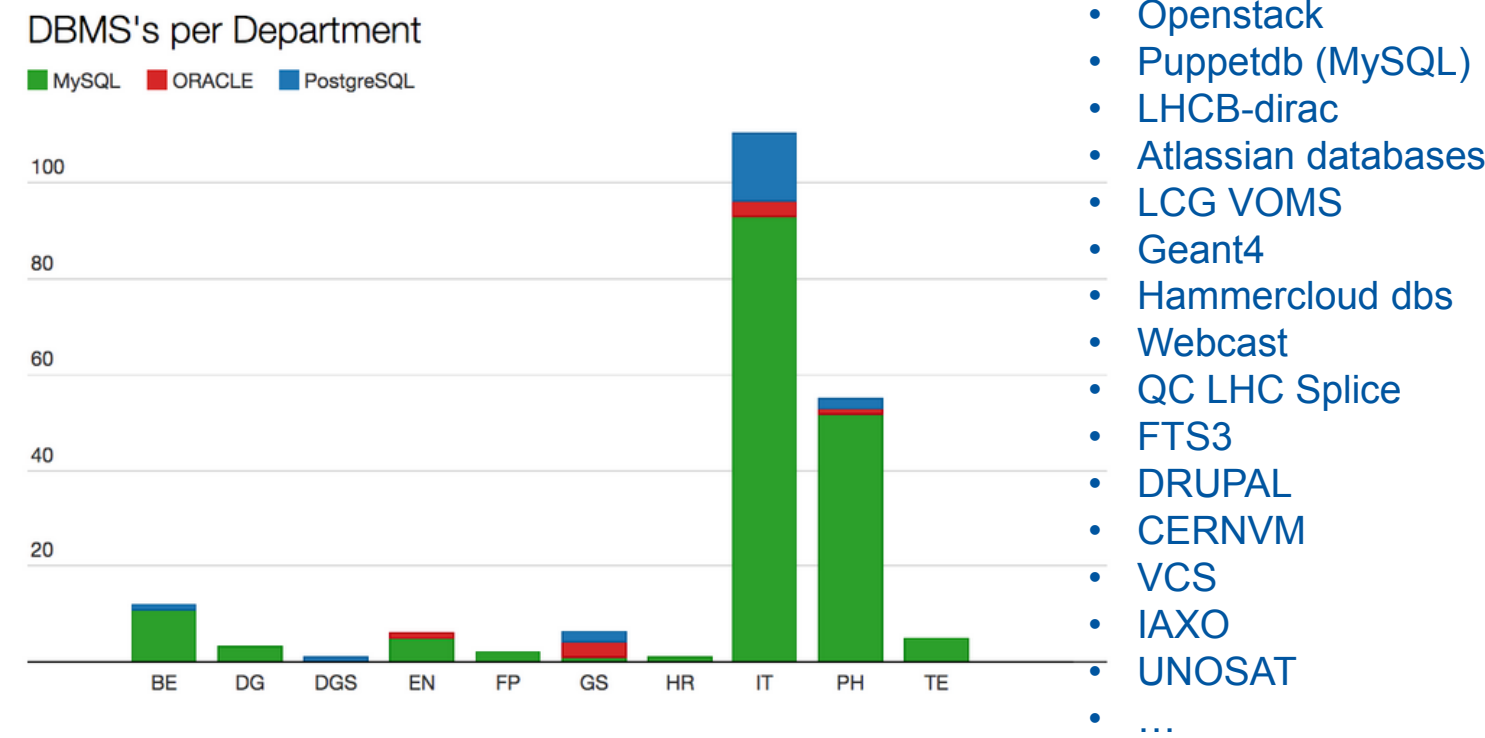
DB on Demand

- Self-service for provisioning, management, backup
 - The number of deployed instances is growing
 - **MySQL** (85%), PostgreSQL (10%) and Oracle (5%)



11

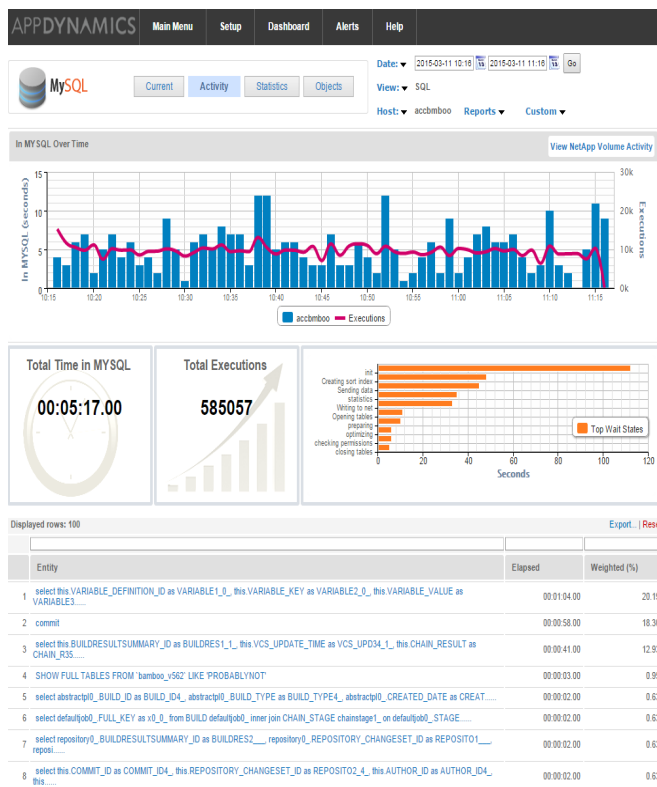
Additional Stats for DBoD



12

DBoD new monitoring tool

- **Appdynamics**
- Allows DBoD users to **troubleshoot** performance problems with a GUI interface



13

- **CERN centric, also going hadoop**
 - <http://indico.cern.ch/event/345619/session/0/contribution/2/material/slides/1.pdf>

~~SDN~~

- Network Engineers are investigating interesting solutions.
- How would you really use P2P? Do you care with today's bandwidth?
- What are your problems?



CDN

- Physics requirements?
 - reduced load on long distance links
 - improved performance for poorly connected sites
- Solutions?
 - cache servers placed in strategic locations of the Internet or LHCONE, a la Google cache or Akamai (i.e. based on peerings rather than geolocalisation)
 - [ipv6] multicast: datasets continuously streamed into multicast groups
 - solution yet to be designed and developed!



11th April 2015

WLCG Networking

26



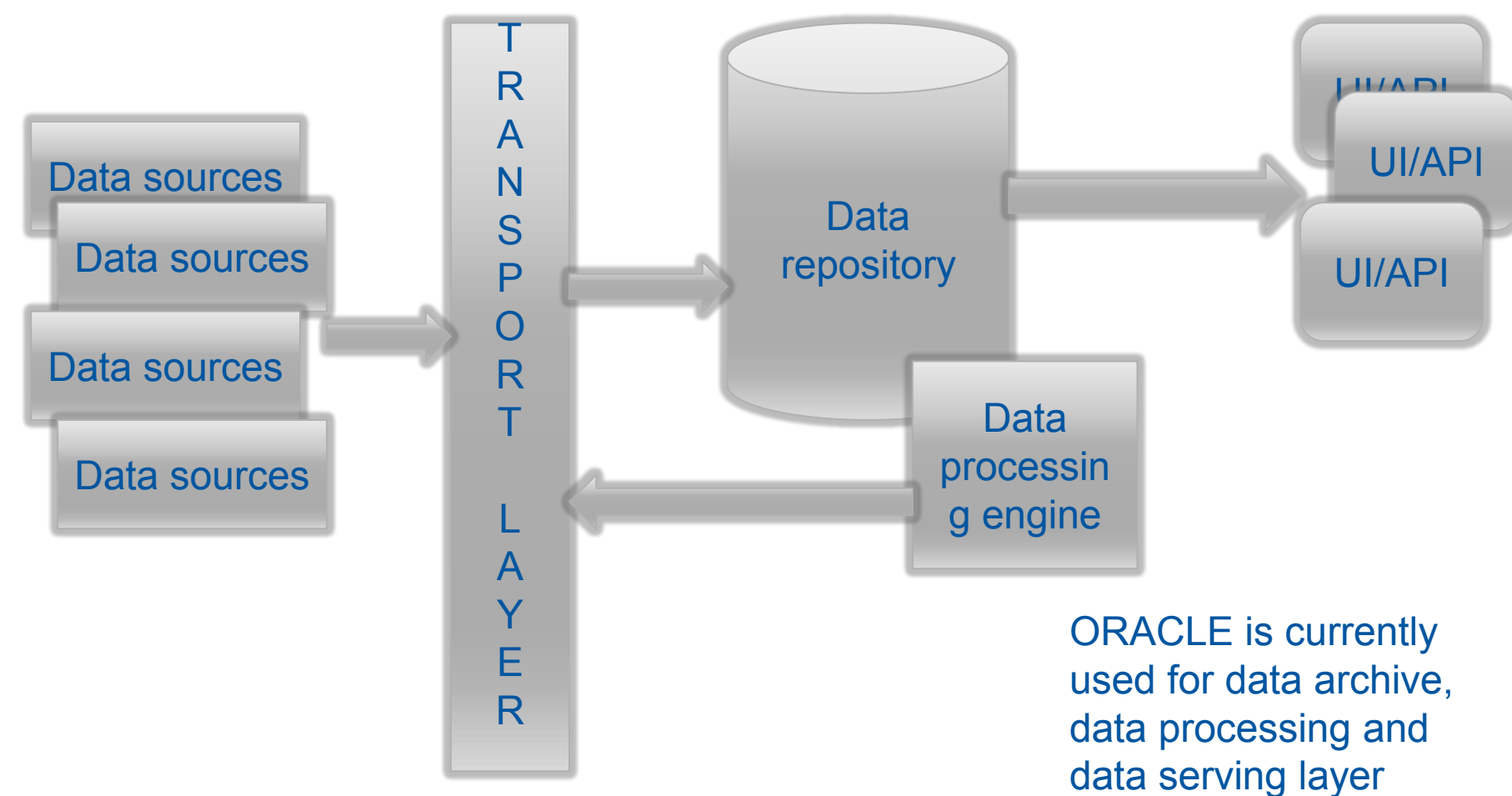
11th April 2015

WLCG Networking

27

- <http://indico.cern.ch/event/345619/session/0/contribution/1/material/slides/1.pdf>

Current architecture



11/04/15

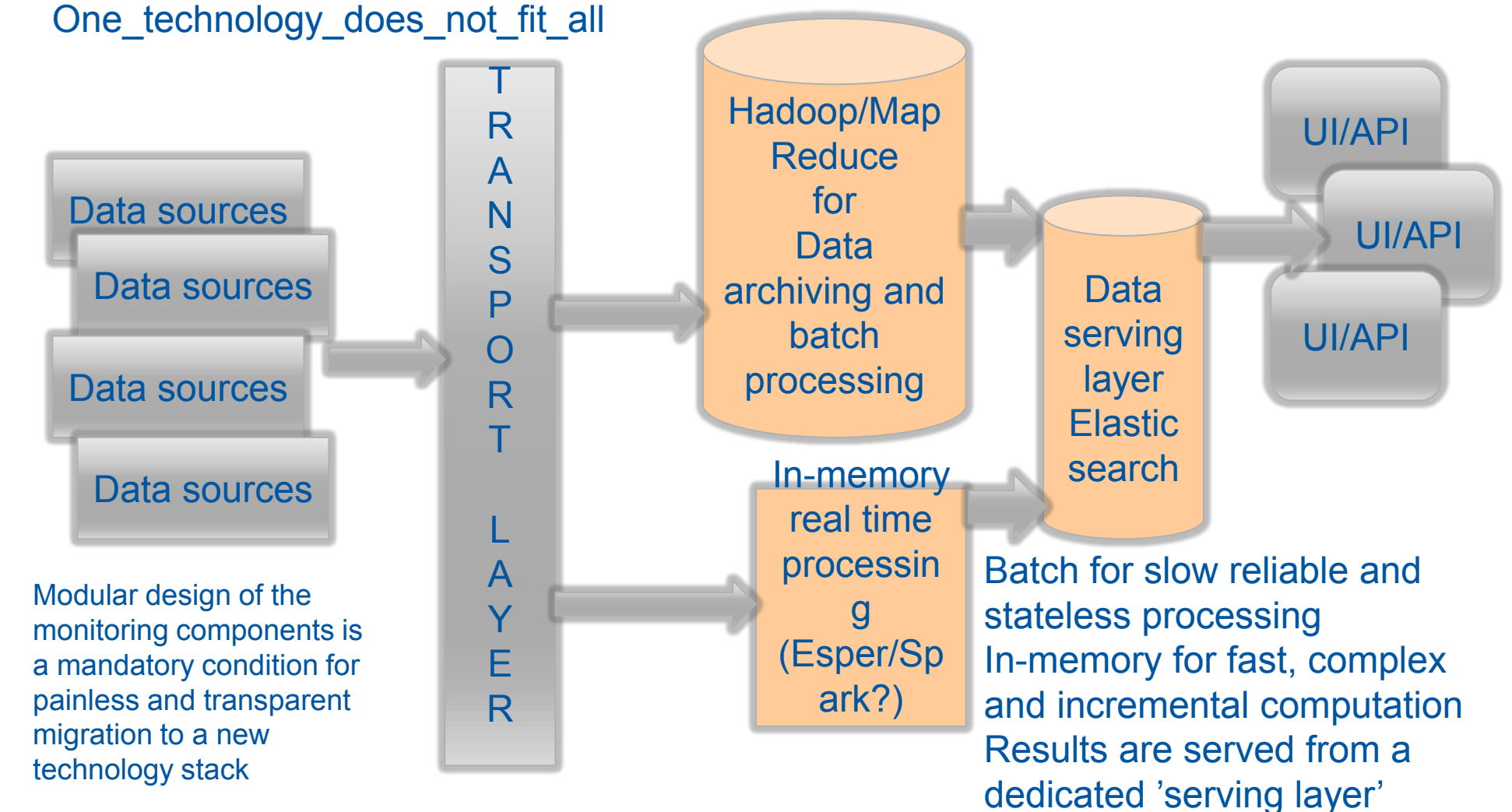
WLCG Workshop, CHEP 2015, Okinawa
Julia Andreeva CERN IT

19

Architecture evolution

Taking inspiration from Lambda architecture

One_technology_does_not_fit_all



11/04/15

WLCG Workshop, CHEP 2015, Okinawa
Julia Andreeva CERN IT

20

- great look at the future: not SQL, but hadoop + elasticsearch

© <http://indico.cern.ch/event/345619/session/0/contribution/5/material/slides/1.pdf>

Batch systems: Two Years of HTCondor at the RAL Tier-1



Ongoing work & future plans

- Integration with private cloud
 - OpenNebula cloud setup at RAL, currently with ~1000 cores
 - Want to ensure any idle capacity is used, so why not run virtualized worker nodes?
 - Want opportunistic usage which doesn't interfere with cloud users
 - Batch system expands into cloud when batch system busy & cloud idle
 - Batch system withdraws from cloud when cloud becomes busy
 - Successfully tested, working on moving this into production
 - See posters at CHEP
- Upgrade worker nodes to SL7
 - Setup SL6 worker node environment in a chroot, run SL6 jobs in the chroot using NAMED_CHROOT functionality in HTCondor
 - Will simplify eventual migration to SL7 – can run both SL6 and SL7 jobs
 - Successfully tested CMS jobs

38



Ongoing work & future plans

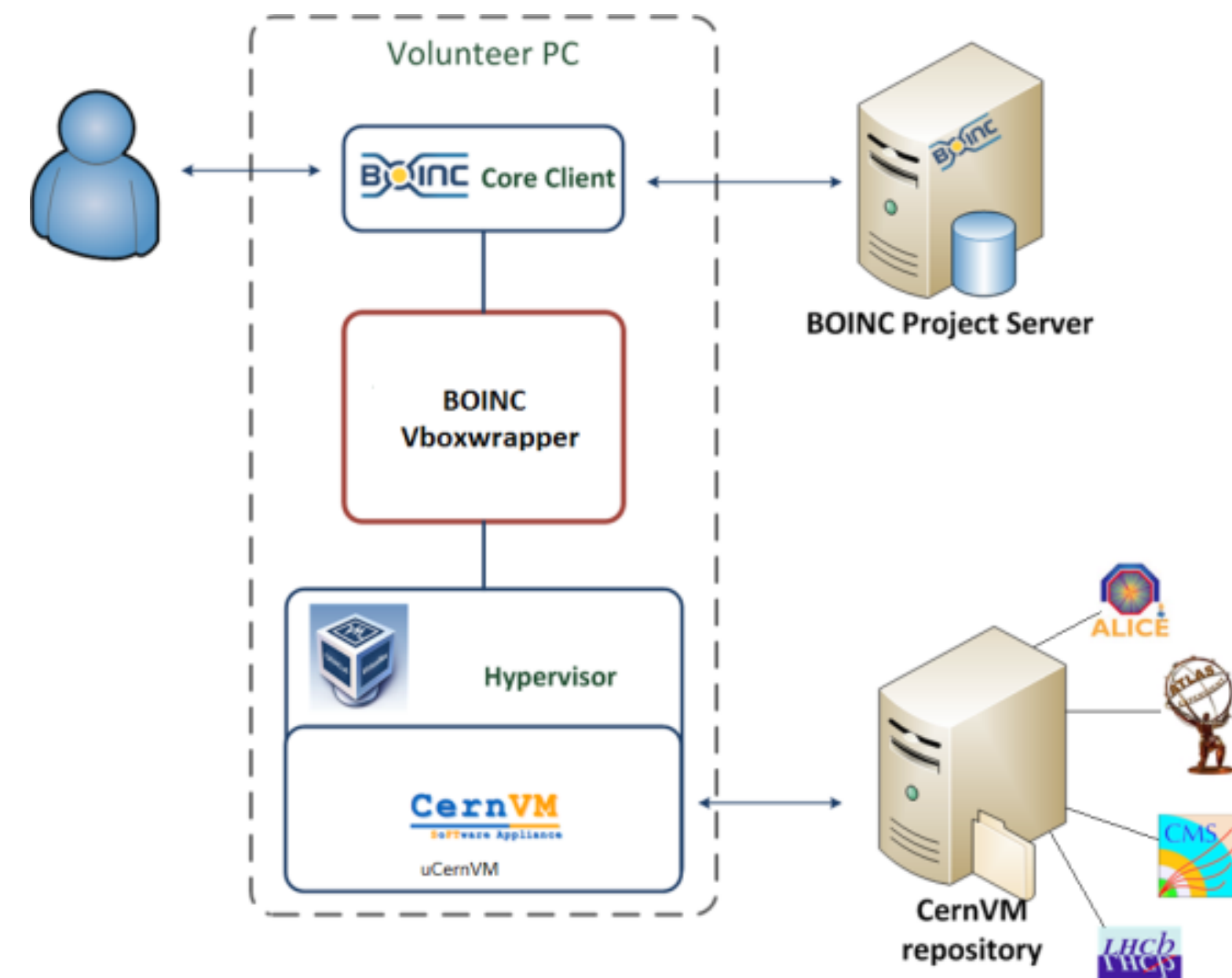
- Simplification of worker nodes
 - Testing use of CVMFS grid.cern.ch for grid middleware
 - 540 packages installed vs 1300 for a normal worker node
 - HTCondor can run jobs:
 - In chroots
 - In filesystem namespaces
 - In PID namespaces
 - In memory cgroups
 - In CPU cgroups
 - Do we really need pool accounts on worker nodes?
 - With the above, one job can't see any processes or files associated with any other jobs on the same worker node, even if the same user
 - Worker nodes and CEs could be much simpler without them!

39

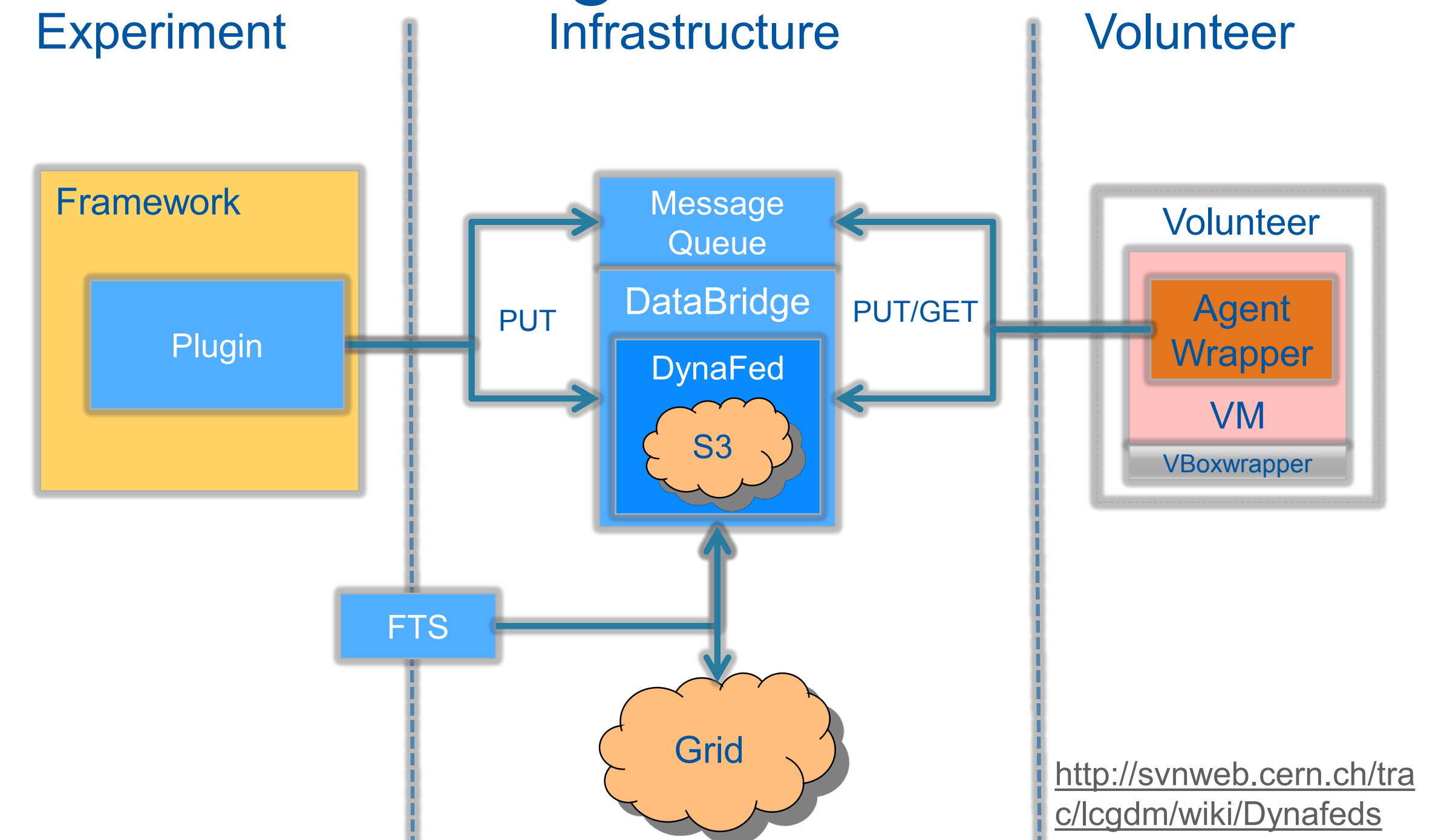
- virtual facility like functionality (ask Andrew Lahiff, CMS person)
 - ◉ <http://indico.cern.ch/event/345619/session/0/contribution/15/material/slides/1.pdf>

Volunteer Computing

BOINC With Virtualization



The DataBridge



- using Boinc and VirtualBox + CernVM

- ◉ <http://indico.cern.ch/event/345619/session/0/contribution/6/material/slides/3.pdf>

Optimisation of operational costs



New Technologies A new type of site

- Cloud
 - WLCG should provide tested and optimised disk images compatible with any cloud infrastructure. Ideally, running a WLCG site should mean running few (1-3) virtual machines with WLCG-provided images configured by a small and well documented configuration file
 - Some of these ideas more expended in the Resource Provisioning presentation
- Containers
 - WLCG should move to a model whereby site admins no longer are required to install local services beyond core cluster functionality;
 - WLCG should invest in containerisation of middleware services to reduce the workload on sites. Keep the barrier of entry for potential opportunistic resource providers at an extreme minimum: provide a grid model but leveraging container technologies such as Docker or Rocket.
 - Some of these ideas have started to be discussed also in the HSF context although not concentrating on grid services.



New Technologies Storage

- Have HTTP-based storage federations
 - Couples with other requests to “stick to industry standards” when evaluating new technologies.
- A big number of sites are looking into CEPH as a storage technology. Among many pros, it enables SSD caching out of the box. This could be a game changer for more efficient WAN transfers. So supporting CEPH as an SE technology could be very beneficial.



A “simple Tier2”

- There is this new concept of a “simple Tier2” but unless we revolutionise our sites there is very little that can be removed. Some suggestions
 - APEL box has been mentioned by several sites as a burden but it is not strictly necessary
 - ARC-CEs (and OSG) publish directly into APEL not clear why all the other CREAM sites couldn't do the same
 - Push new sites towards ARC-CE which is simpler and more robust
 - Push new sites towards Htcondor as community is building up
 - Alternatively SLURM or if they can afford it UGE or latest LSF
 - Re-evaluate the need of an heavy weight BDII, mostly used for service discovery and getting few unreliable numbers in Rebus.
 - Are all the lines it publishes really needed?
 - Service in itself is light weight **if something fills the values for you. And YAIM is fading.**
 - Keep up the work to reduce the number of storage protocols

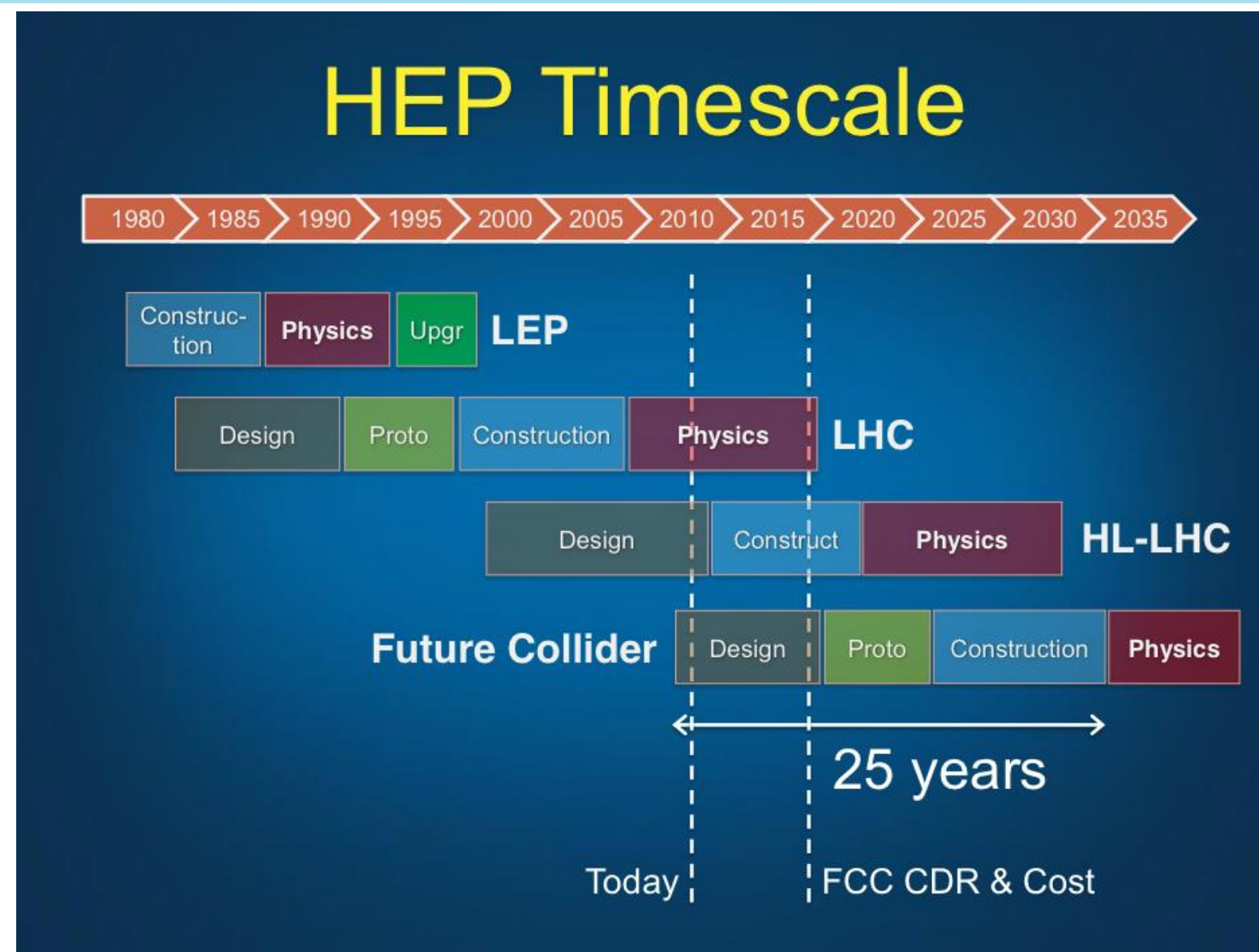


▪ WLCG simple T2: ARC-CE, HTCondor

▪ Re-evaluate the need of an heavy weight BDII, mostly used for service discovery and getting few unreliable numbers in Rebus.

- <http://indico.cern.ch/event/345619/session/0/contribution/3/material/slides/0.pdf>

Computing at the HL-LHC timescale: introduction



Trends in HEP computing

- Distributed computing is here to stay
 - Actually we had it 30 years ago, and seriously 15-20 years ago
- Ideal general purpose computing (x86 + Linux may be close to the end)
 - May be more effective to specialise
 - GPU and other specialised farms
 - HPC machines
 - Commodity processors ("x86", ARM, etc)
 - Used for different purposes – lose flexibility but may gain significantly in cost



23 March 2015

Ian Bird; FCC Week

6



23 March 2015

Ian Bird; FCC Week

11

Trends – software

- Recognizing the need to re-engineer HEP software
 - New architectures, parallelism everywhere, vectorisation, data structures, etc.
- Set up HEP Software Foundation (HSF)
 - Community wide – buy in from major labs, experiments, projects
 - Goals:
 - Address rapidly growing needs for simulation, reconstruction and analysis of current and future HEP experiments,
 - Promote the maintenance and development of common software projects and components for use in current and future HEP experiments,
 - Enable the emergence of new projects that aim to adapt to new technologies, improve the performance, provide innovative capabilities or reduce the maintenance effort,
 - Enable potential new collaborators to become involved,
 - Identify priorities and roadmaps,
 - Promote collaboration with other scientific and software domains.

- CERN Council decided that HL-LHC makes proposal to be part of ESFRI
 - ◉ European Strategy Forum for Research Infrastructures, was submitted end of March
- Asking general questions of where to go, how to evolve, worth a read!
- <http://indico.cern.ch/event/345619/session/1/contribution/23/material/slides/1.pdf>

Evolution of facilities

- Today we have LHC/WLCG as the computing facility
- Recognise that between now and FCC, we have potentially many international facilities/collaborations involving global HEP community
 - Bearing in mind we have possibly many international or global HEP challenges: Neutrino facility, ILC, CLIC, FCC and others as well as large experiments such as Belle-II that ask to use "WLCG"
 - And not forgetting the possible commonalities with related projects (SKA, LSST, CTA, etc) where facilities may be heavily shared
- What is the process to build on our working infrastructure to evolve towards HL-LHC, FCC, etc. serving the needs of these facilities and learning from them?
- How should WLCG position itself to help build a common global infrastructure that evolves through these coming facilities?



23 March 2015

Ian Bird; FCC Week

12



23 March 2015

Ian Bird; FCC Week

13

Evolution of structure

- Distinguish between infrastructure and high level tools
- We need to continue to build and evolve the basic global HEP (+others) computing infrastructure
 - Networks, AAA, security, policies, basic compute and data infrastructure and services, operational support, training, etc.
 - This part MUST be common across HEP and co-existing science
 - This part must also be continually evolving and adapting with technology advances
- Need a common repository/library of proven and used middleware and tools
 - A way to help re-use of high and low level tools that help an experiment build a computing system to make use of the infrastructure
 - The proto-HSF today could be a seed of this
- We must try and make this a real common effort and remove a lot of today's duplication of solutions
 - While retaining the ability and agility to innovate
 - The cost of continuing to support unnecessary duplication is too high



Scale Numbers

	HLT Output	Events per year	RAW per Event	RAW data per year
Run1	600Hz	3.6B	0.7MB	2.5PB
Run2	1kHz	5B	1.0MB	6PB
Run3	1kHz	5B	1.2MB	7.2PB
Run4	5kHz	25B	2.5MB	75PB

- Assuming an (optimistic) physics beam time of 6M seconds per year
 - However, this is the target for HL-LHC to collect 300fb⁻¹ per year
- What will the relationship between RAW data and derived data be?

4

How hard does it get?

- Event Generation
 - Not intrinsically harder at high luminosities, however better generators and studying rare processes will mean using more cycles; volume increase to scale with events
- Simulation
 - Main scaling of simulation per event is with energy (so ~constant in Runs 2, 3, 4, ...); however, more data needs more simulation to accompany it, so volume increases
- Digitisation
 - More or less linear with pile-up as background minimum bias events are layer on top of signal events; more simulation → more digitisation
- Reconstruction
 - Definitely *very hard* at high pile up; scaling is naively $\mu!$ (factorial) for tracking; certainly the biggest challenge faced in software; combines with volume increases
- Analysis
 - Most likely linear with data volumes, but analysis can already be i/o bound; thus i/o becomes a serious problem; need to optimise across huge range of workloads

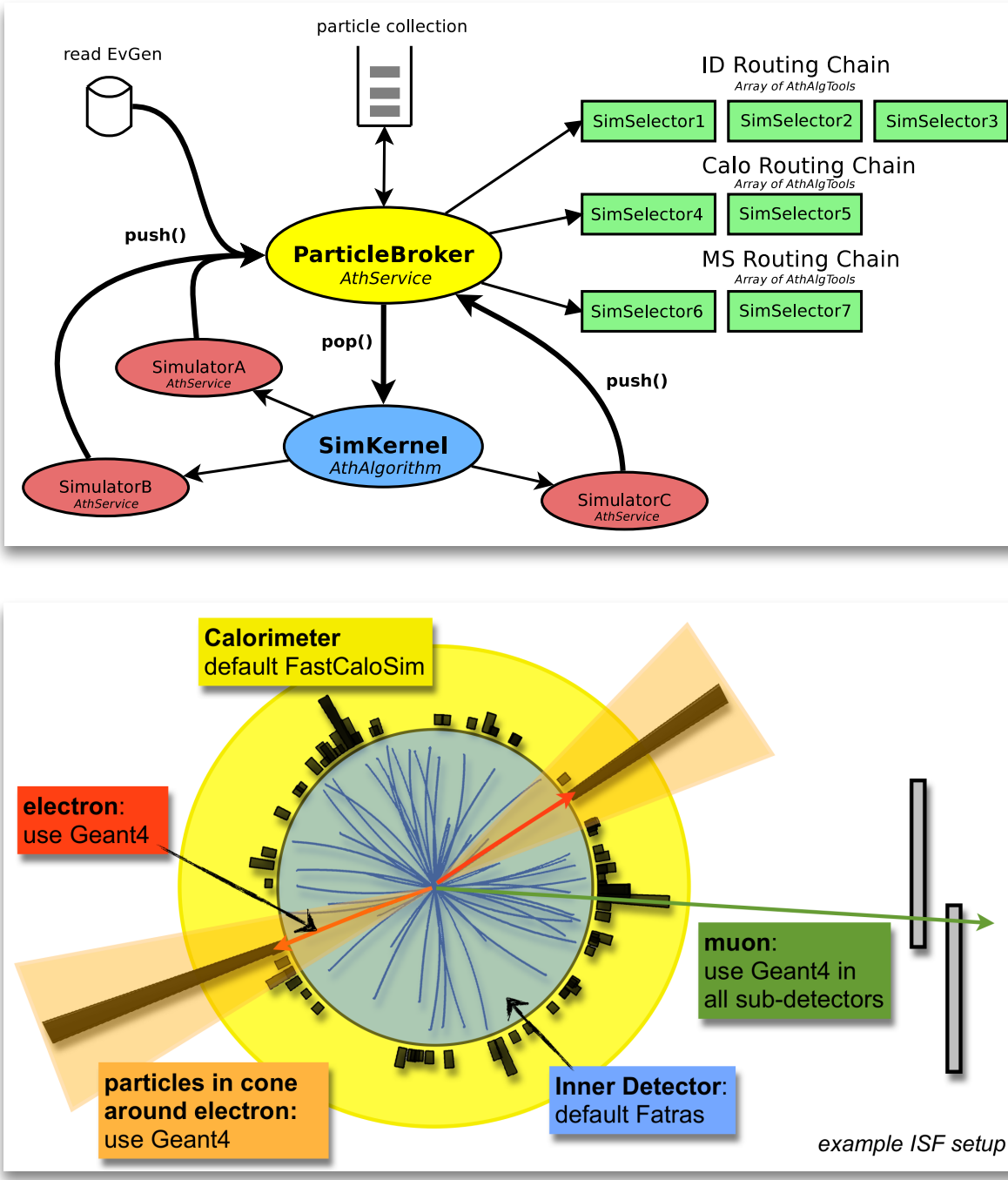
5

- good overview, worth a read
- <http://indico.cern.ch/event/345619/session/1/contribution/9/1/material/slides/0.pdf>

Integrated Simulation Framework

- Single framework for simulation
- Simulation engines act like services
- Choose engine based on particle type and region of interest
- Mix simulation types within a single event
- Full potential realised when combined with fast digitisation and reconstruction

Tracker	Calo.	Muons	speedup
full	fast	full	~20
fast	fast	fast/full	>100
Rol guided fast/full			~100



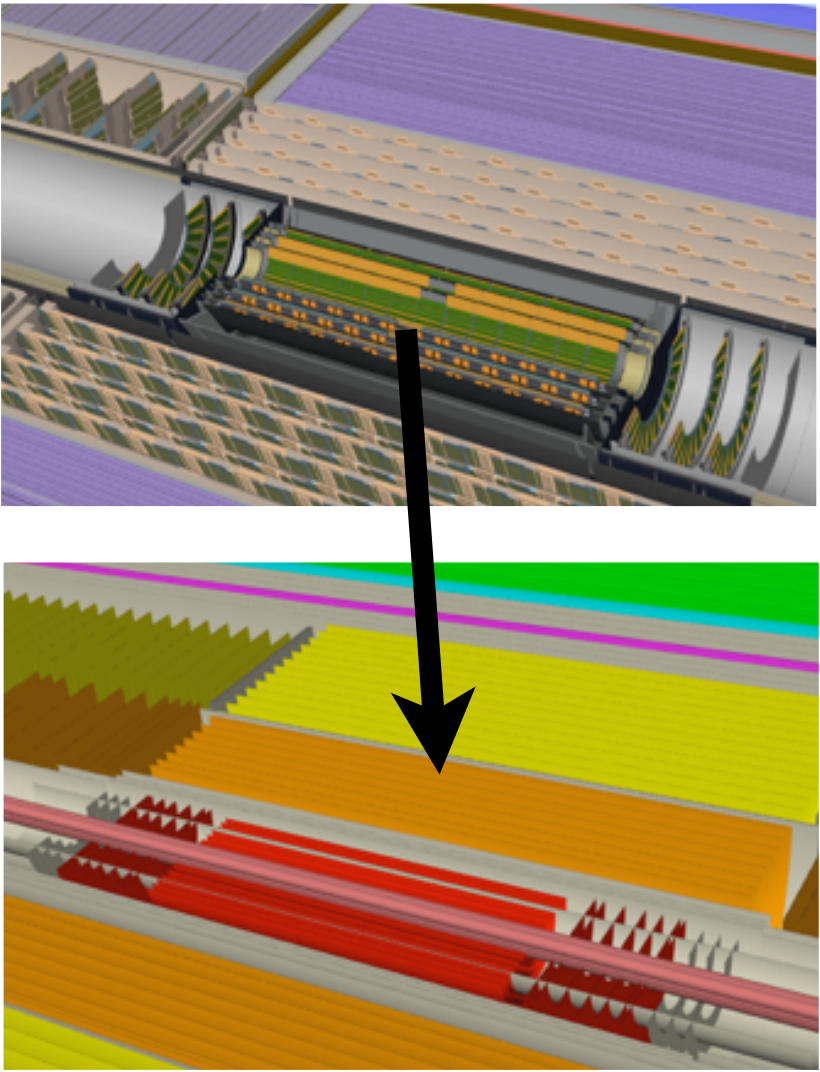
9

Andy Salzburger, Elmar Ritsch

Fast Simulation: FATRAS

- ATLAS has 2 geometry systems (not special)
 - Full model used in Geant4 with 4.8M placed volumes
- Reconstruction model for fast tracking
 - reduced complexity
 - material projected onto surfaces
- Fast extrapolation engine
 - embedded navigation replaces voxialization
- Fatras simulation engine
 - re-uses track reconstruction infrastructure
 - combined with particle stack and fast physics processes
 - optionally: fast digitisation codes

ATLAS	G4	tracking	ratio
crossed volumes in tracker	474	95	x5
time in SI2K sec	19.1	2.3	x8.4



Andy Salzburger, Markus Elsing

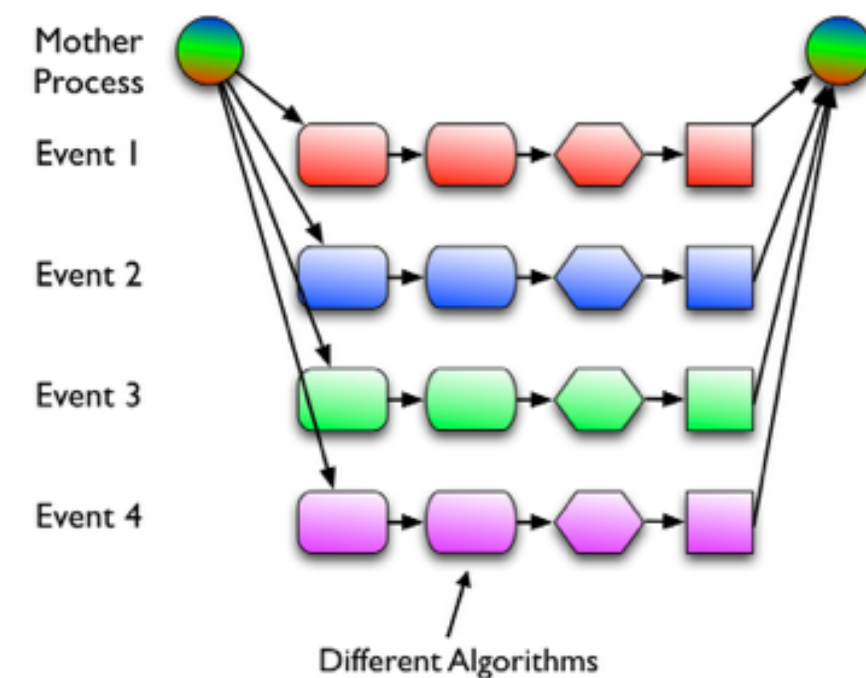
10

- good overview, worth a read
- <http://indico.cern.ch/event/345619/session/1/contribution/9/1/material/slides/0.pdf>

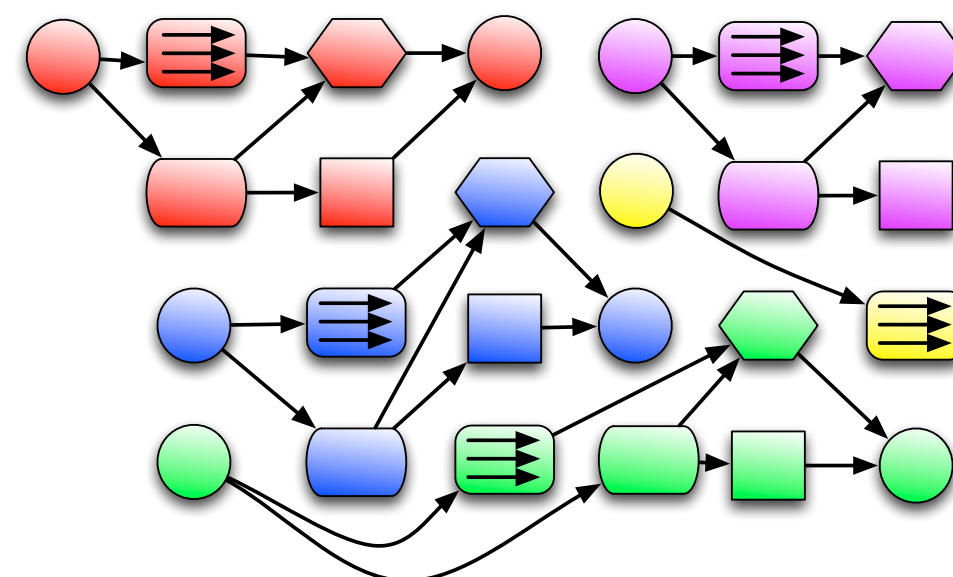
New Framework: GaudiHive

- Memory constraints, especially on non-Xeon server architectures make reducing memory footprint imperative
 - High luminosity and hard tracking conditions only increase this pressure
- Need to move to a multi-threading framework (beyond AthenaMP)
 - Memory savings can be huge as all heap memory is shared
 - However, a more difficult programming model as threads can interfere with each other: data races and deadlocks
- Development to introduce parallelism into the Gaudi framework used by ATLAS and LHCb
- Take advantage of parallelism between algorithms and across multiple events
- Scheduler is data flow driver, but control flows can also be given (important for online)

13



Run2 AthenaMP multi-processing: Each worker uses a separate process, but read-only memory pages are shared



Run3 multi-threaded reconstruction: Colours represent different events, shapes different algorithms; all one process running multiple threads

Computing in 10 years...

- This is very hard to predict, but
 - Certainly need custodial storage for RAW data
 - Large quantities of disk for online data
 - Fronted by smart caches of fast storage?
 - (The trick is not to cache what we just used, but what we are just going to use — hinted pre-caching via PanDA, ARC)
 - Will need to manage carefully volumes of derived and simulated data
 - Archive to tape more aggressively than in the past
- Storage services and compute services could increasingly decouple
 - Fast, smart networks funnelling data where it's needed
- Allows for easier use of heterogeneous resources
 - HPC, spot priced clouds, BOINC, ...
- Classic WLCG sites will probably get bigger and more efficient
 - Evolution towards wider scientific remit (HPC/HTC convergence) as well as reducing costs and maintaining expertise
 - Smaller resources migrate to lightweight stacks — BOINC clients?

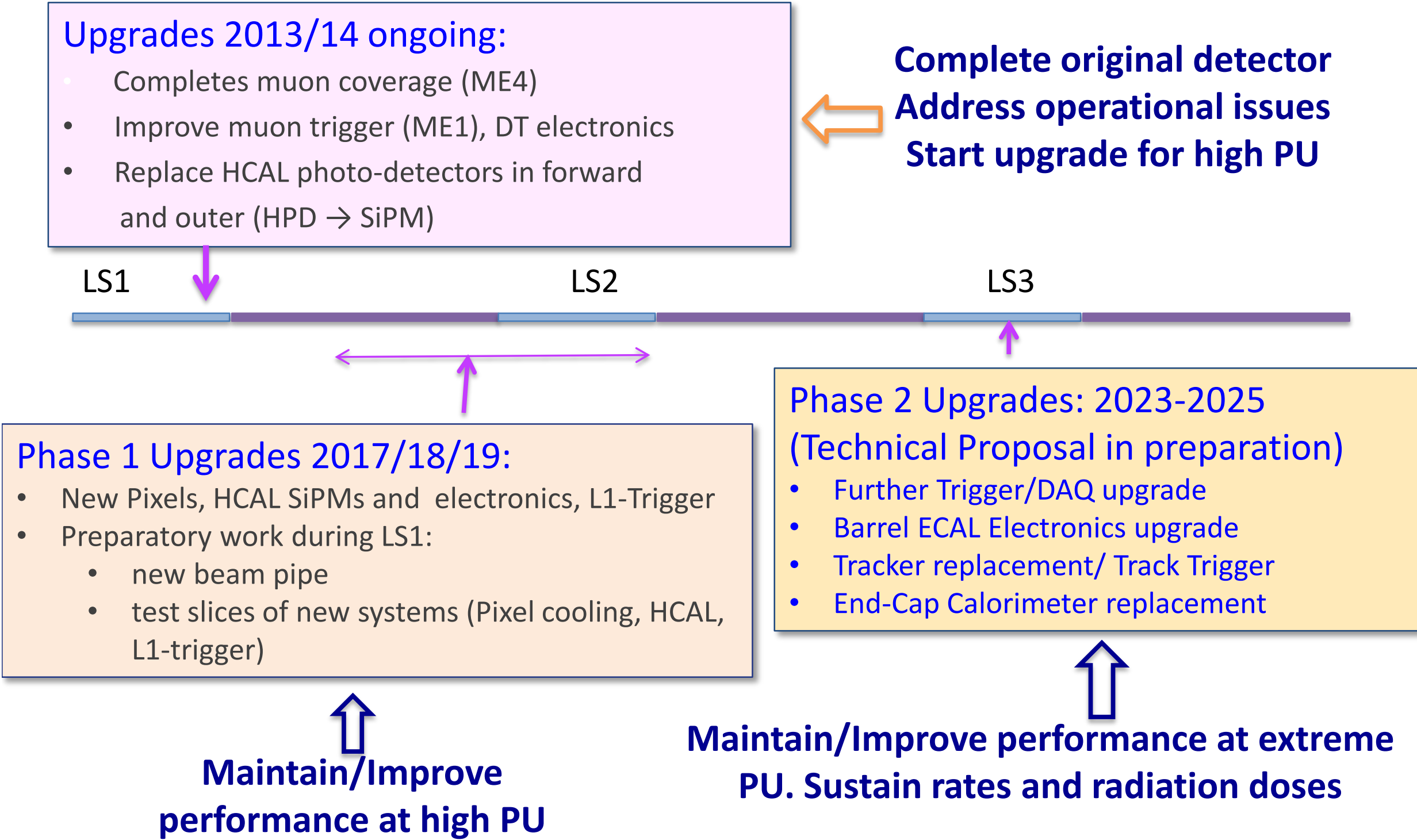
16

▪ good overview, worth a read

▪ <http://indico.cern.ch/event/345619/session/1/contribution/9/1/material/slides/0.pdf>

Computing at the HL-LHC timescale: CMS

CMS Upgrade Strategy - Overview



WLHG@Okinawa

2

Estimating Resource Needs

- CMS is planning for 5-7.5kHz of data in CMS Phase-II. In this scenario CMS would collect **25B-37B** raw events per year
 - Estimating from the current software and using the upgrade simulation we see that each of these events is more complicated to reconstruct and larger than the events we will collect in 2015

	Detector	Pile-up (Ave./crossing)	Reconstruction time (Ratio to Run 2)	AOD size (Ratio to Run 2)
Run3	Phase 1	50	4	1.4
	Phase-II	140	20	3.7
Run4	Phase-II	200	45	5.4

WLHG@Okinawa

6

▪ <http://indico.cern.ch/event/345619/session/1/contribution/28/0/material/slides/1.pdf>

Computing at the HL-LHC timescale: CMS

HL-LHC focused R+D Activities and ideas (Reports during CHEP on many areas)

- Towards using heterogeneous many-core resources
 - We have ported the CMS track reconstruction to the Intel Phi
- Trigger on many-core or specialized resources
- Simulation (G4 or beyond) on heterogeneous resources
- Resource scheduling, I/O, event processing frameworks for heterogeneous resources, parallel processing models
 - Developing a computing model simulation to study optimizing data access
- Improving tracking and other algorithms at high pile-up
 - Continuous improvement activities
- Evaluating improved cost/performance of using specialized centers for dedicated workflows

WLHG@Okinawa

15

HL-LHC focused R+D Activities (Reports during CHEP on many areas)

- Evaluating “big data” data reduction and selection techniques for I/O intensive analysis procedures
 - Efficient physics object skimming/tagging would also promote re-use of selection criteria
- Developing tools and infrastructure (profilers, dev. tools, etc.)
 - Prototyping container technology (e.g., DOCKER) for CMSSW
- Evaluating metrics for performance per power use, in addition to simply raw performance
- Investigating data analytics techniques
 - Ideas for next generation data popularity

WLHG@Okinawa

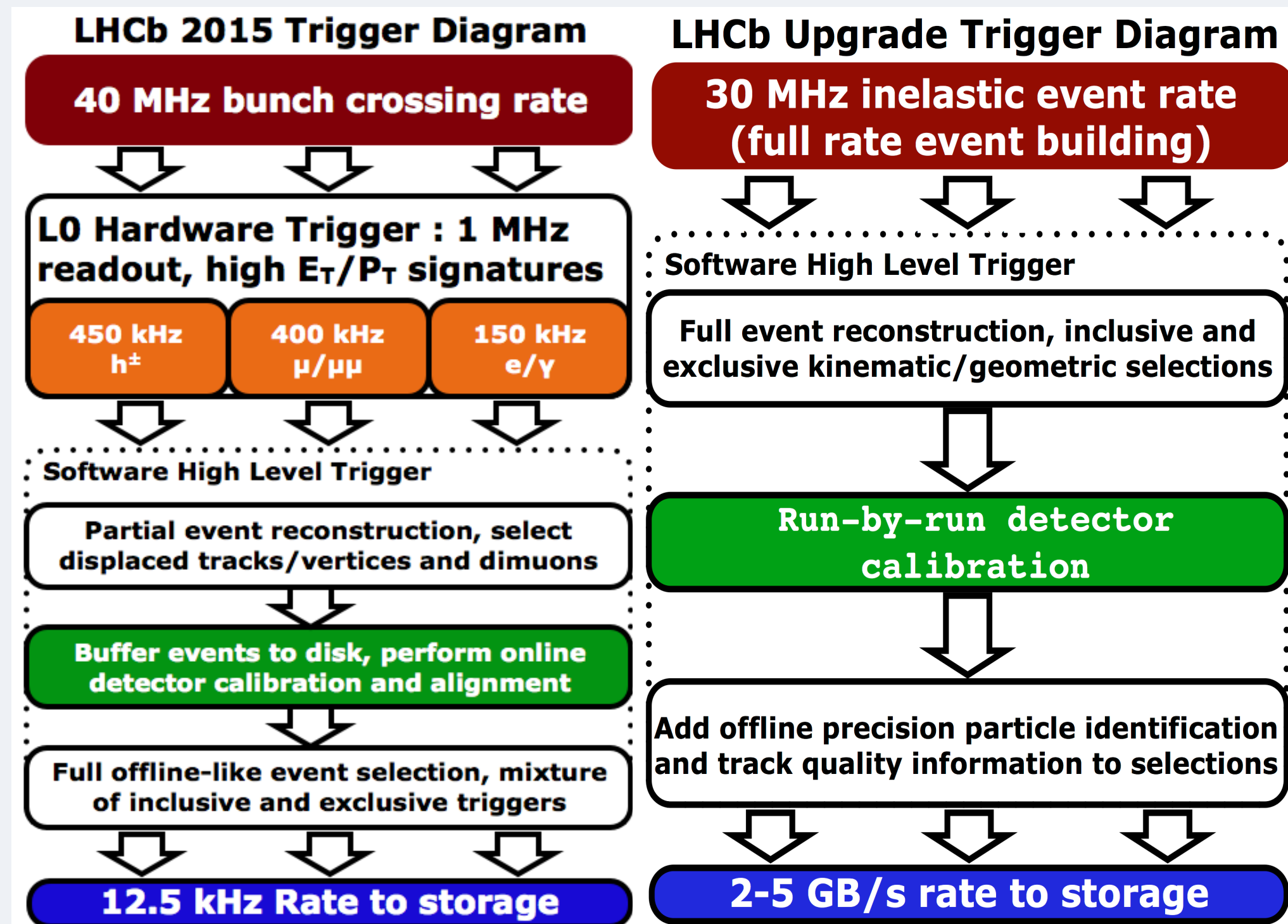
16

▪ <http://indico.cern.ch/event/345619/session/1/contribution/28/0/material/slides/1.pdf>

Computing at the HL-LHC timescale: LHCb



Trigger now and at the upgrade



5



Brainstorming: The Game Has Changed

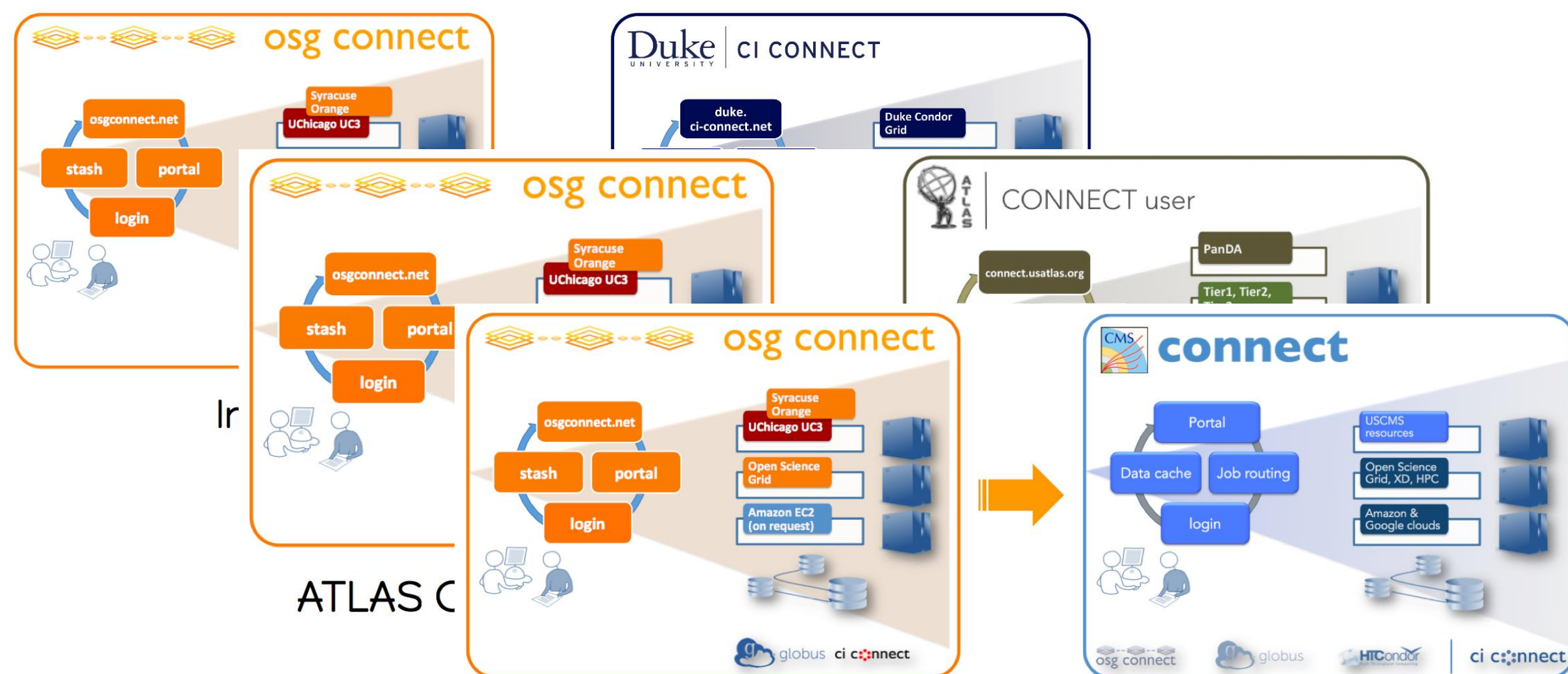
- Trigger is no longer selecting events, but classifying them
 - Write out what bandwidth and offline resources allow, but everything written out will be analysed
- In many exclusive analyses, interested only in the decay tree of the triggering signal
 - So write out only the interesting part of the event, not the whole event
 - Turbo Stream idea, being tested in Run 2
 - ☆ 2.5kHz of signal, ~5kB/event, going directly to analysis
 - ☆ x10 reduction in event size
- If all events are interesting offline, current model of stripping no longer applies
 - Streaming more relevant, but how many streams?
 - Is direct access to individual events more relevant?
 - ☆ Needs event index, and R&D on efficient access to single events



10

▪ <http://indico.cern.ch/event/345619/session/1/contribution/28/2/material/slides/0.pdf>

OSG as a Service



CMS Connect: Analysis service for the US CMS Collaboration

Slides by Rob Gardner

18

Continued focus on the “Long Tail of Science”

- “OSG will be considered successful if we not only keep up with the growing LHC needs but also expand to enhance the computing throughput to a broad spectrum of scientists at a variety of scales, from individual users at a single campus to multi-institutional experiments”
- <http://indico.cern.ch/event/345619/session/2/contribution/18/material/slides/1.pdf>

Parting Words

- OSG is ready for LHC Run 2 data
- embrace the “long tail” communities and disciplines while at the same time keeping focus during LHC Run 2
- future success is about minimizing obstacles
 - Providing gateways to abstract complexities
 - OSG Connect portal, login, data services
 - ongoing Galaxy project (bioinformatics portal)
 - ongoing partnership with EGI competence centers doing similar work (MoBrain and ELIXIR)
- travel and time zone issues can make collaboration difficult, the OSG is committed to a continued partnership with WLCG

20

Hardware technology trends

Market Dominance

Only a few large companies are dominating the various components markets

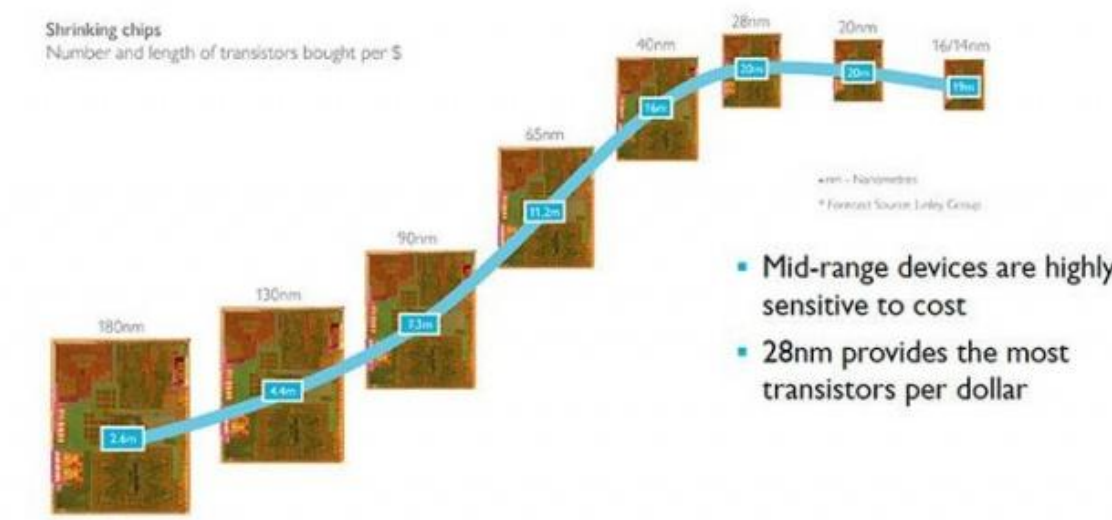
Processors	INTEL, Qualcomm, Samsung, AMD
Graphics	INTEL, Nvidia, AMD
Hard Disk Drives	Western Digital, Seagate, Toshiba
DRAM memory	Samsung, SK Hynix, Micron
NAND Flash memory	Samsung, Toshiba, SanDisk, Micron, Hynix, INTEL
Solid State Disks	Samsung, INTEL, SanDisk, Toshiba, Micron
FPGA	Xilinx, Altera (currently being bought by INTEL)
Tape Storage	HP, Fuji, IBM, SpectraLogic ORACLE, IBM

RoI Return-on-Investment is the keyword
Few companies capable of large scale investments, majority fabless companies
Favour evolutionary (adiabatic) changes of technology
Clear bias against 'disruptive' new technologies
(memristor, holographic storage, DNA storage, quantum computing, non-volatile memory, etc.)

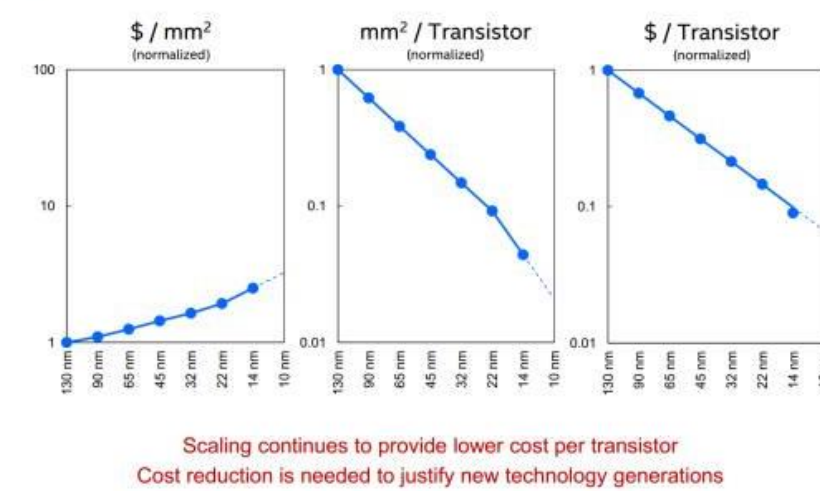
e.g. Yearly revenues: Samsung 209 B\$ INTEL 56 B\$

Processor Technology, Moore's Law

28nm: Optimal Balance of Cost and Power for 2015 Devices

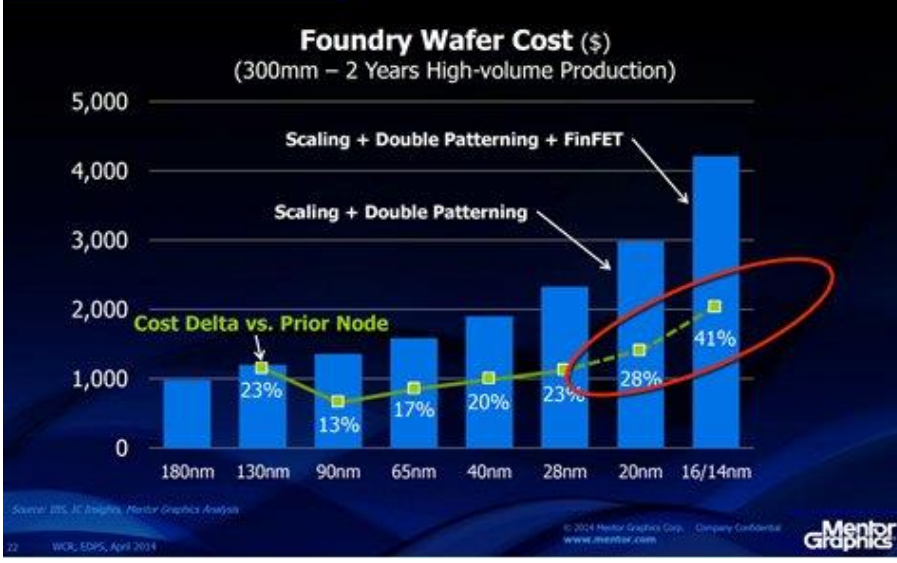


(EP1) Moore's Law Challenges Below 10nm: Technology, Design and Economic Implications



INTEL claims to overcome this up to the 10nm node scale

Traditionally, Cost-per-Wafer Increases 15-20% at Each New Technology Node



Quite some discussion in 2014 about the end of Moore's Law

Moore's Law is about the production cost of transistors not about the sales cost of processors

- HEP is ~\$15M out of a total \$52B market
- <http://indico.cern.ch/event/345619/session/1/contribution/10/material/slides/1.pdf>

Hardware technology trends

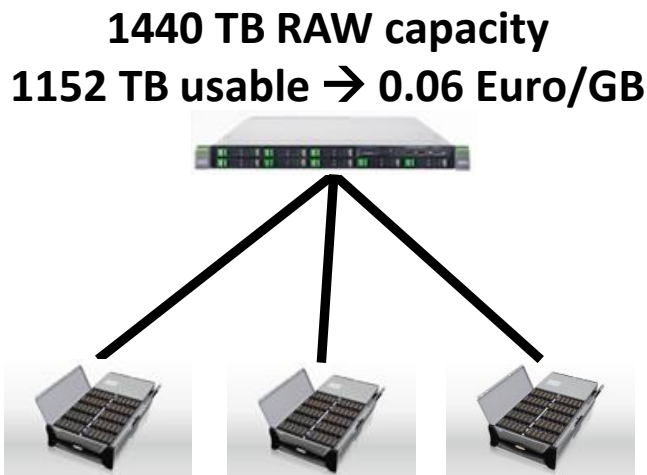
Back-of-an-Envelope Calculations, storage component savings

CERN disk server: CPU server with SAS attached JBOD array



Example: 'improve' the storage costs by a factor 3:

4 TB server disk ~0.05 Euro/GB → 8 TB SMR ~0.03 Euro/GB (low-end desktop 6 TB)
Dual 24-bay disk tray → three 60-bay disk trays per frontend
RAID0 / data replica → Erasure code, data increase by 1.25 instead of 2



This improves the space costs but reduces considerable the IO capabilities. But how much IO do we actually need ?
(Application, data management, data distribution dependent)
Much more tuning between application and hardware needed.....

Redefine our notion of storage space
→ Storage space plus performance

different IO architecture based on Seagate Kinetic
object drive model or the HGST Open Ethernet drive

Split
MC+processing facilities -- analysis facilities

FLAPE
Flash+Tape

Summary

- Semiconductor Component and end-user markets are stabilizing.
Saturation effects seen nearly everywhere, moving to 'replacement' markets
- Very few companies dominating the market: technology evolution , not revolution
- Moore's Law validity being debated. 3D technology helps.
Expect still continuous price/performance improvements, but lower levels
- Server market is small compared to the consumer market, stable and highly profitable
Market --> high prices. Microservers show in principle potential, but currently overrated
- Way to improve price/performance beyond the technology --> architecture
- Should not talk about disk, SSD or tape but rather storage units (space+performance)
- There will be processing and storage technologies in 2025 and most likely not too different from today, but estimating the cost is pretty difficult.
So.. You will get what you get (equal or rather lower budget than today).....

- HEP is ~\$15M out of a total \$52B market
- <http://indico.cern.ch/event/345619/session/1/contribution/10/material/slides/1.pdf>

CERN Disk Storage Overview

	AFS	CASTOR	EOS	Ceph	NFS	CERNBox
Raw Capacity	3 PB	20 PB	140 PB	4 PB	200 TB	1.1 PB
Data Stored	390 TB	86 PB (tape)	27 PB	170 TB	36 TB	35 TB
Files Stored	2.7 B	300 M	284 M	77 M (obj)	120 M	14 M

AFS is CERN's linux home directory service

CASTOR & EOS are mainly used for the physics use case (Data Analysis and DAQ)

Ceph is our storage backend for images and volumes in OpenStack

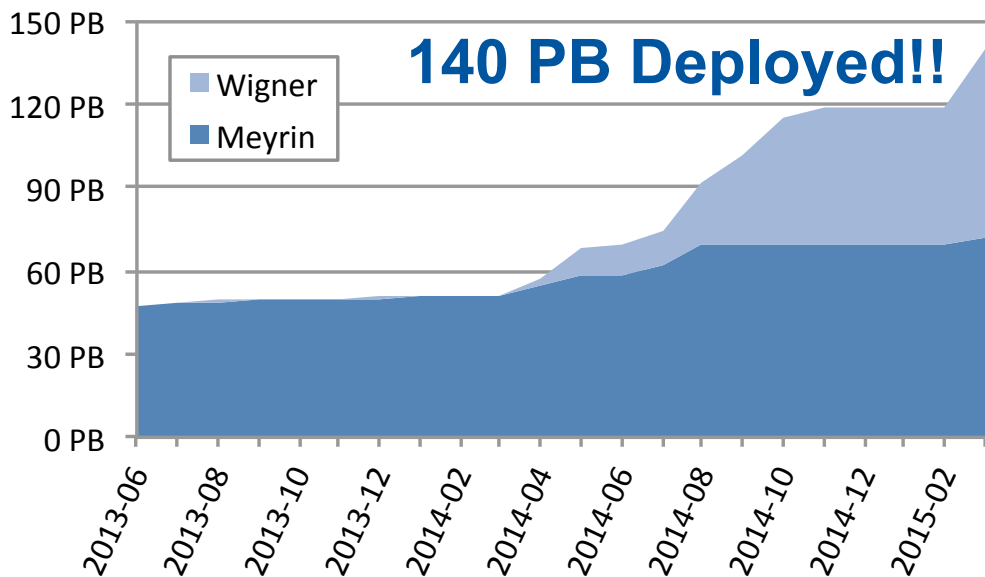
NFS is mainly used by engineering application

CERNBox is our file synchronisation service based on OwnCloud+EOS



EOS @ Wigner

EOS Installed Raw Disk Capacity

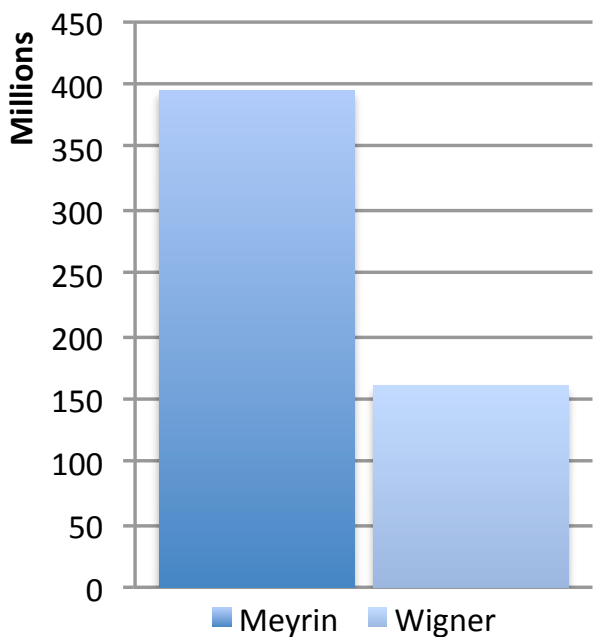


EOS is now optimised for managing efficiently data in different computer centre providing to our user a single site view

And in the future it will be possible to specify adhoc scheduling policy based on the namespace location

Easy to add to the system other locations

EOS Replicas Distribution

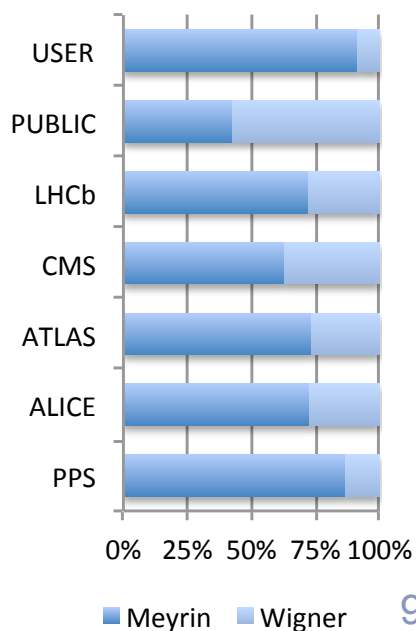


The latest hardware delivery (Mar 2015) balanced the capacity installed in the 2 computer centres (~50% ~50%)

Experiments replicas are not yet spread equally between the 2 geolocation

Geo-balancing need to be activated and tuned to avoid filling the network links to Wigner

Replicas Distribution in % per Experiment

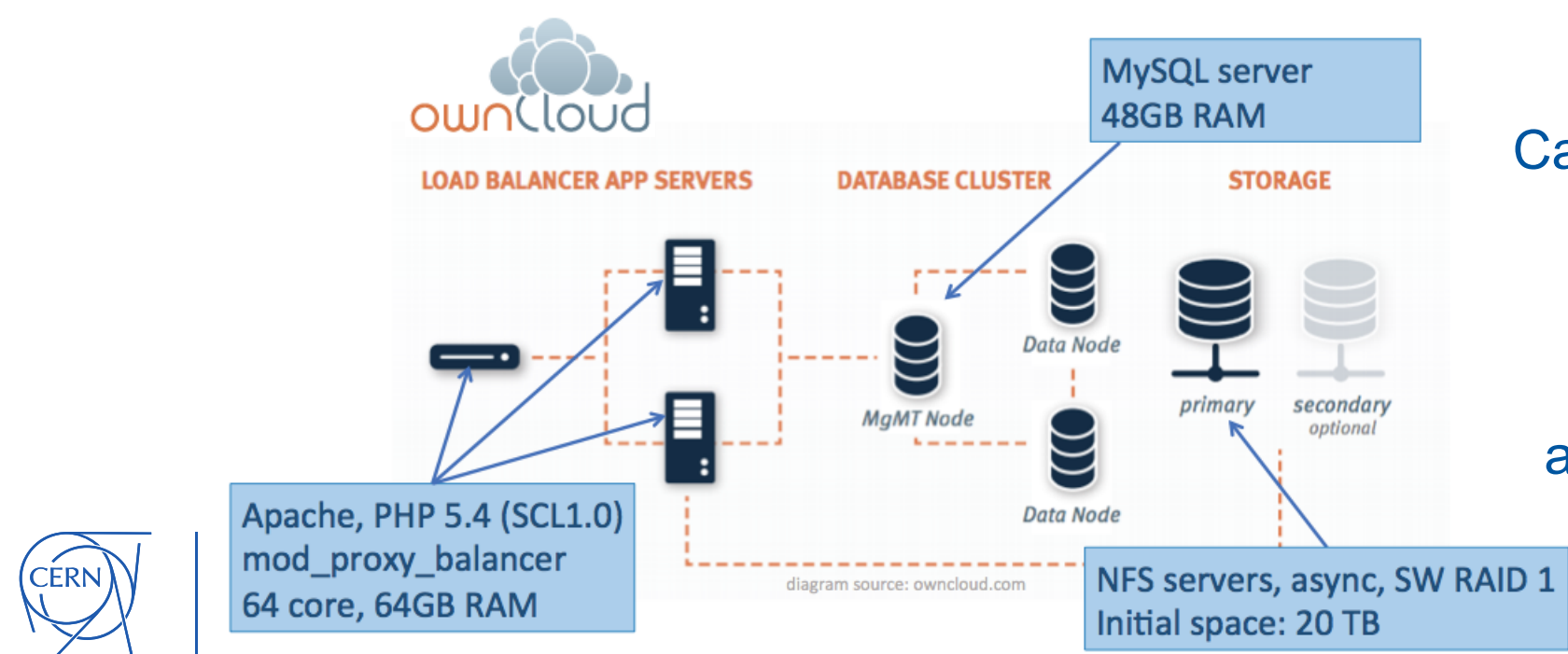


- Lots of Ceph
- <http://indico.cern.ch/event/345619/session/1/contribution/17/material/slides/0.pdf>

New Storage Technologies

Why CERNBox?

- Competitive alternative to Dropbox for CERN users
 - users were using dropbox not only for sharing their pictures
 - SLAs: data availability and confidentiality
 - Archival and Back-up policies
 - Offline Data Access and Data sync across devices
 - Easy way to share files and folders with colleagues
- We started ownCloud evaluation and build prototype service



Can we integrate sync & share functionality with our main users workflow?

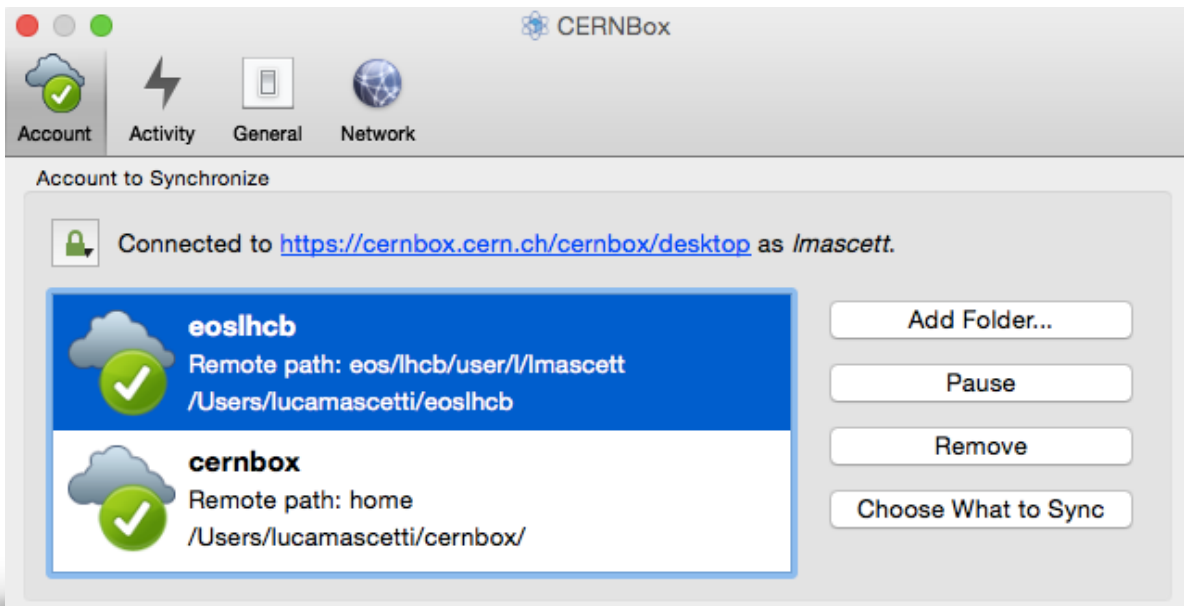
And being able to directly access the underlying data?

11

Available soon...

- Direct access to EOSUSER (and not only...)
 - not only own cloud sync client
 - xroot, fuse, http/WebDAV
- Access to Physics Data
 - synchronise experiment's data
- Direct access from lxplus and batch
 - sync from your laptop and run!
 - sync results back

```
[lmascett@lxplus2015 ~]#  
[lmascett@lxplus2015 ~]# df -H -t fuse  
Filesystem      Size  Used Avail Use% Mounted on  
eosuser         506T   70T  437T   14% /eos/user  
eosatlas        36P    17P   20P   45% /eos/atlas  
eosalice        20P    11P   8.5P   57% /eos/alice  
eoscms          28P    14P   15P   49% /eos/cms  
eoslhcb         13P    7.6P   4.6P   63% /eos/lhcb  
eospublic       16P    5.8P   11P   36% /eos/public  
[lmascett@lxplus2015 ~]#  
[lmascett@lxplus2015 ~]# ls -lc /eos/user/l/lmascett/  
total 6644  
drwx----- 1 lmascett c3      5 Dec 10 15:58 CERN  
drwx----- 1 lmascett c3      0 Jan 26 18:18 debug  
drwx----- 1 lmascett c3      0 Dec 11 09:43 download  
drwx----- 1 lmascett c3      0 Oct 31 18:24 pdf  
drwx----- 1 lmascett c3      1 Dec 11 09:44 personal  
drwx----- 1 lmascett c3      8 Dec 10 12:11 pictures
```



15

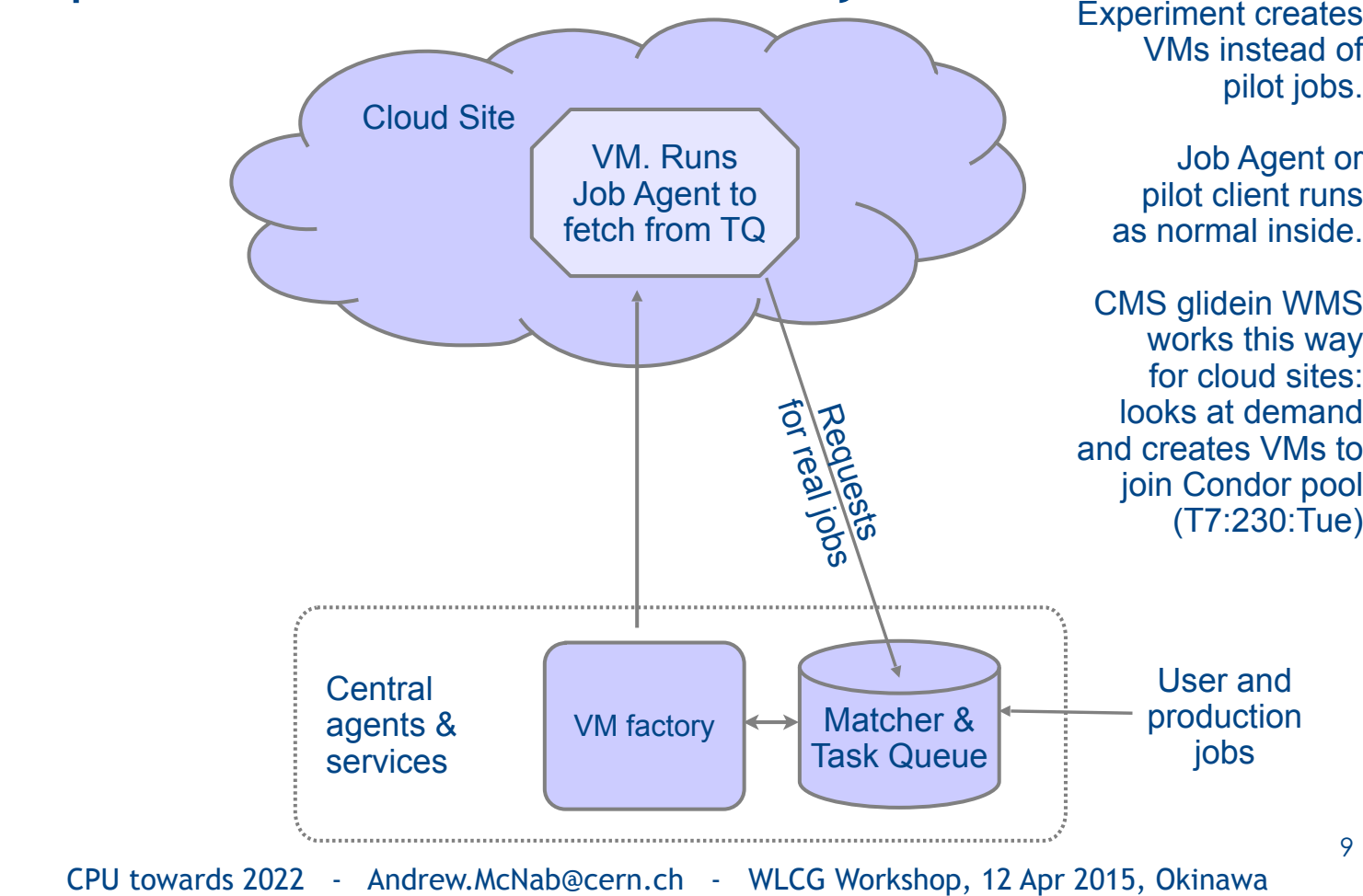
- Lots of Ceph
- <http://indico.cern.ch/event/345619/session/1/contribution/17/material/slides/0.pdf>

Resource Provisioning: Overview

Resource Groups	CernVM	Commercial Clouds
<ul style="list-style-type: none">• WLCG Tier 0/1/2• HLT Farms• Commercial Providers• Volunteer Computing• HPC• Parameter space contains 160 possibilities<ul style="list-style-type: none">– 5 resource groups, 8 areas, 4 VOs– Need consolidation of solutions• The cloud paradigm<ul style="list-style-type: none">– The adoption of generic solutions<ul style="list-style-type: none">• Reducing our total cost of ownership	<ul style="list-style-type: none">• The OS via CVMFS<ul style="list-style-type: none">– HTTP replication of a reference file system<ul style="list-style-type: none">• Stratum 0• Why?<ul style="list-style-type: none">– Because CVMFS is already a requirement<ul style="list-style-type: none">• Reduces the overhead of distributed image management<ul style="list-style-type: none">– Manage version control centrally• CernVM as a common requirement<ul style="list-style-type: none">– Parameter space reduction<ul style="list-style-type: none">• 20 possibilities => 1 CernVM<ul style="list-style-type: none">– 5 resource groups, 4 VOs• Availability becomes an infrastructure issue– Potentially 20 different contextualizations<ul style="list-style-type: none">• Responsibility of the VO• The goal is to start a CernVM-based instance<ul style="list-style-type: none">– There are exceptions where this is not possible	<ul style="list-style-type: none">• Helix Nebula<ul style="list-style-type: none">– A public-private partnership<ul style="list-style-type: none">• Between research organizations and IT industry• Microsoft Azure Pilot<ul style="list-style-type: none">– Preliminary discussions with CERN OpenLab• Amazon<ul style="list-style-type: none">– BNL RACF for ATLAS and CMS– With new Scientific Computing group at AWS• Deutsche Börse Cloud Exchange AG<ul style="list-style-type: none">– Beta testing platform– Will go live beginning of May• PICSE<ul style="list-style-type: none">– Procurement Innovation for Cloud Services in Europe• European Science Cloud Pilot<ul style="list-style-type: none">– Pre-Commercial Procurement (PCP) proposal<ul style="list-style-type: none">• Buyers group public organizations that are members of the WLCG collaboration

▪ <http://indico.cern.ch/event/345619/session/1/contribution/11/0/material/slides/1.pdf>

Experiment creates VMs directly?



- Following the CHEP 2013 paper:
 - *“The Vacuum model can be defined as a scenario in which virtual machines are created and contextualized for experiments by the resource provider itself. The contextualization procedures are supplied in advance by the experiments and launch clients within the virtual machines to obtain work from the experiments’ central queue of tasks.”*
(“Running jobs in the vacuum”, A McNab et al 2014 J. Phys.: Conf. Ser. 513 032065)
 - a loosely coupled, late binding approach in the spirit of pilot frameworks
- For the experiments, VMs appear by “spontaneous production in the vacuum”
 - Like virtual particles in the physical vacuum: they appear, potentially interact, and then disappear
- CernVM-FS and pilot frameworks mean a small `user_data` file and a small CernVM image is all the site needs to create a VM
 - Experiments can provide a template to create the site-specific `user_data`

- Three VM Lifecycle Managers that implement the Vacuum model
- Vac is a standalone daemon run on each worker node machine to create its VMs
 - At Manchester, Oxford, Lancaster, Birmingham
- Vcycle manages VMs on IaaS Clouds like OpenStack
 - Run by the site, by the experiment, or by regional groups like GridPP
 - Resources at CERN (LHCb), Imperial (ATLAS, CMS, LHCb), IN2P3(LHCb)
 - Vcycle instances running at CERN, Manchester, Lancaster
 - Vac/Vcycle talk T7:271:Mon
- HTCondor Vacuum manages VMs on HTCondor batch systems
 - Injects jobs which create VMs; VM jobs can coexist with normal jobs
 - Running at STFC RAL. See T7:450:Mon
- All make very similar assumptions about how the VMs behave
 - The same ATLAS, CMS, LHCb, GridPP DIRAC VMs working in production with all three managers

CPU towards 2022 - Andrew.McNab@cern.ch - WLCG Workshop, 12 Apr 2015, Okinawa

Why HSF?

- The challenges are large, but why do we think the HSF can help?
- It's a mechanism to facilitate coordination and common efforts in HEP software and computing.
- We need to exploit all the expertise available in our community, and outside it, to meet these challenges.
- The affordable way to do it, is collaboratively.

8

Agreed HSF Goals

- Share expertise
- Raise awareness of existing software and solutions
- Catalyze new common projects, create an incubator
- Promote commonality and collaboration in new developments to make the most of limited resources
- Aid developers and users in discovering, using and sustaining common software
- Support training career development for software and computing specialists
- Provide a framework for attracting effort and support to S&C common projects
- Provide a structure for the community to set priorities and goals for the work
- Facilitate wider connections; while the HSF is a HEP community effort, it should be open enough to form the basis for collaboration with other sciences

12

▪ <http://indico.cern.ch/event/345619/session/1/contribution/13/material/slides/0.pdf>